TRIP TIMING AND SCHEDULING PREFERENCES

Programme de recherche ERA-NET SURPRICE

Convention de subvention 10 MT -SURPRICE - 2-CVS-070-2010

Rapport final

2 novembre 2012

Ecole Normale Supérieure de Cachan (ENS Cachan)

Technical University of Denmark (DTU)

Royal Institute of Technology (KTH)

Un projet du PREDIT 4, GO 6 : Politiques de transports

Dans le cadre du programme européen ERA-NET, ENT 17 SURPRICE.

Ecole Normale Supérieure de Cachan (ENS Cachan)
Département d'Economie et Gestion
61 avenue du président Wilson
94320 Cachan, FRANCE.
Contact : Professeur André de Palma, Courriel: andre.depalma@ens-cachan.fr

Technical University of Denmark (DTU)
Department of Transport
Bygningstorvet 116B
2800 Kgs. Lyngby, DENMARK.
Contact : Professeur Mogens Fosgerau, Courriel: mf@transport.dtu.dk

Royal Institute of Technology (KTH)
Centre for Transport Studies
Teknikringen 14
10044 Stockholm, SWEDEN.
Contact : Dr. Anders Karlström, Courriel: anders.karlstrom@kth.se

Equipe:

Prof. André de Palma, ENS Cachan, France (Coordinateur)

Prof. Mogens Fosgerau, DTU, Denmark

Dr. Anders Karlström, KTH, Stockholm, Sweden

TABLE DES MATIERES

I. REMERCIEMENTS

Nous adressons nos remerciements à Mr. Gérard Brun pour l'intérêt particulier qu'il porte à nos travaux. Ses commentaires, remarques et suggestions nous ont permis d'avancer sur de nombreux points.

Nous remercions aussi tous nos collègues pour leurs remarques au cours de nombreuses discussions.

Nous remercions particulièrement Robin Lindsey pour ses commentaires et suggestions.

II. RAPPEL DU CONTEXTE

L'objet de la recherche est de contribuer au corpus théorique et scientifique de l'analyse des coûts de déshorage dans les trois situations suivantes : la prise en compte de l'interaction des choix individuels au sein d'un couple, la problématique des coûts endogènes, et la prise en compte du problème de stationnement. A l'exception de la problématique du stationnement, les sujets abordés sont nouveaux.

La démarche utilisée est analytique dans le sens où elle s'intéresse au développement de cadres théoriques d'analyse reposant sur des modélisations mathématiques de ces problèmes économiques, i.e. comment les politiques de tarification et de règlementation doivent-elles tenir compte des coûts de déshorage dans les trois situations envisagées ?

Ce rapport final présente trois contributions : la première porte sur l'analyse du trip timing et des préférences de déshorage, la deuxième sur la tarification du stationnement comme substitut à la tarification routière dans le cadre d'un modèle dynamique avec congestion et la dernière sur les files d'attentes aléatoires en présence d'usagers averses au risque.

III. TRIP TIMING - PREFERENCE POUR LES CHOIX DES HEURES DE DEPART

Mogens Fosgerau, Technical University of Denmark et Center for Transport Studies, Suède

André de Palma, Ecole Normale Supérieure de Cachan et Centre d'économie de la Sorbonne. CES, France

Anders Karlstrom, KTH Royal Institute of Technology et Centre for Transport Studies, Suède

Ken Small, University of California, Irvine, USA

RESUME

Ce rapport résume les résultats du projet intitulé SURPRICE: Trip timing and scheduling preferences. Dans ce projet nous nous focalisons sur l'importance de choix de l'heure de déplacement comme une des causes de la congestion. Il est important de reconnaître que le choix de l'heure de départ revient à l'usager et que la congestion se développe quand nombreux voyageurs décident de voyager au même moment. Pour la mise au point et l'évaluation d'un schéma de tarification, nous devons prendre en compte explicitement l'évolution de la distribution de l'heure de départ particulièrement si on applique une tarification modulaire. Sans prise en compte de choix de l'heure de déplacement, l'identification et l'évaluation d'une part important des bénéfices de tarification sera impossible et le schéma choisi ne pourra pas être optimal. Ce projet répond à 4 questions fondamentales dans la procédure de mise au point et d'évaluation des schémas de tarification des usagers : l'interaction entre la forme urbain et le choix de l'heure de déplacement, mécanismes alternatives de tarification pour l'utilisation optimale des capacités, la nature des préférences en choix de l'heure de déplacement et des modèles dynamiques d'affectation du trafic et la nature de l'équilibre dans ces modèles.

Trip timing and scheduling preferences

Mogens Fosgerau[1]
André de Palma[2]
Anders Karlstrom[3]
Kenneth Small[4]

June 6, 2012

Abstract

This note summarizes the results from the project SURPRICE: Trip timing and scheduling preferences. The general emphasis of this project is the importance of trip timing as a cause of congestion. It is important to recognize that departure time is a choice of travellers and that congestion arises because many travelers choose to travel at the same time. The design and evaluation of pricing schemes should explicitly take changes in departure time patterns into account, in particular with time-varying charges. Failure to take trip timing into account will lead to failure in identifying important benefits and will lead to less efficient pricing schemes. This project has addressed four fundamental questions of vital interest to the design and evaluation of road user charging (RUC) schemes: The interaction of urban structure with trip timing, alternatives to pricing mechanisms for allocating capacity, the nature of scheduling preferences and models for dynamic assignment and nature of equilibrium.

Keywords: Road user charges; Congestion pricing; Road pricing; Trip timing; Scheduling preferences; Dynamic assignment models

[1] Technical University of Denmark, Denmark & Centre for Transport Studies, Sweden, mf@transport.dtu.dk

[2] Ecole Normale Supérieure de Cachan, Centre d'Économie de la Sorbonne (CES), Cachan, France, andre.depalma@ens-cachan.fr

[3] Centre for Transport Studies, KTH Royal Institute of Technology, Sweden, anders.karlstrom@kth.se

[4] Department of Economics, University of California, Irvine, USA, ksmall@uci.edu

1 Introduction

The congestion of the morning rush hour is a prominent feature of urban life. Aggregate congestion delay is a significant burden on industrialized economies and much attention has been given to policy measures, notably pricing, that can reduce congestion.

This note summarizes the results from the project SURPRICE: Trip timing and scheduling preferences. The general emphasis of this project is the importance of trip timing as a cause of congestion. It is important to recognize that departure time is a choice of travellers and that congestion arises because many travellers choose to travel at the same time. The design and evaluation of pricing schemes should explicitly take changes in departure time patterns into account, in particular with time-varying charges. Failure to take trip timing into account will lead to failure in identifying important benefits and will lead to less efficient pricing schemes.

Most current traffic models are notoriously poor in incorporating trip timing. Much remains to be done in making models both sufficiently realistic in this respect and at the same time sufficiently simple for application. There are furthermore many remaining challenges at the conceptual level.

This project has addressed four fundamental questions of vital interest to the design and evaluation of road user charging (RUC) schemes.

- The interaction of urban structure with trip timing
- Alternatives to pricing - mechanisms for allocating capacity
- The nature of scheduling preferences
- Models for dynamic assignment and nature of equilibrium

It is first important to understand and to be able to model trip timing in an urban context. The challenge is that travellers, located at various distances from their destination, face different scheduling constraints. This has an impact on their choices and hence on the resulting equilibrium, such that the shape of the rush hour depends on the spatial structure of a city. Hence spatial structure must be taken into account when designing tolls. This project is the first to achieve this goal in an analytical model.

Second, the view of congestion that recognizes equilibrium in the timing of trips opens for alternatives to pricing. These could be a further development of the high occupancy/toll (HOT) lanes and high-occupancy vehicle (HOV) lanes that have been implemented in the US. Various ways can be devised of assigning a share of capacity to designated groups of vehicles for some period of time

during the peak. Such mechanisms have the potential to reduce congestion without pricing and without capacity expansion. Under bottleneck queueing they can be designed such that no traveller is worse off than before the scheme. This is important for acceptability.

Third, in order to account for trip timing choices of travellers, it is necessary to improve our current understanding of scheduling preferences. In particular, while the state-of-the-art so far has treated scheduling preferences as exogenous, it is reasonable to think that they are in fact endogenous: a commuter's preferred arrival time at work depends essentially on when everybody else arrive at work. Taking this into account may change forecasts and assessment of the optimal toll.

Fourth, for applications it is important to understand the nature of equilibrium in dynamic assignment models. In this project we will address the theory based on concepts of equilibrium, and its relation to applied dynamic assignment models, used in forecasting.

This project is the first to address these questions. The project has been carried out in collaboration between leading researchers within transport economics from Denmark, France, Sweden, and the US.

2 The interaction of urban structure with trip timing

The research that is summarized in this section has been published in Journal of Urban Economics in a paper with the title "Congestion in a city with a central bottleneck" (Fosgerau and de Palma, 2012). This paper presents a model that integrates two prominent features of urban congestion, focusing on the exemplary case of the morning commute. The first feature is that congestion is a dynamic phenomenon in the sense that congestion at one time of day affects conditions later in the day through the persistence of queues. The second feature is that trip origins are spatially distributed. We analyze how these features interact in a city with a central bottleneck and provide results concerning optimal pricing.

2.1 Background

The dynamics of congestion were analyzed in the seminal Vickrey (1969) bottleneck model (see also Arnott et al., 1993), which captures the essence of congestion dynamics in a simple and tractable way. Travellers are viewed as having scheduling preferences concerning the timing of trips that have to pass the bottleneck. The analysis concerns equilibrium in the traveller choice of departure time.

The Vickrey (1969) analysis of congestion, however, essentially ignores space.

Using the notation of the current paper, travellers are depicted as travelling some distance c (measured in time units) until they reach a bottleneck at time a. They exit the bottleneck to arrive at the destination at time t. They have scheduling preferences, always preferring to depart later and always preferring to arrive earlier. The Vickrey (1969) scheduling preferences can be expressed by the scheduling cost $\alpha \cdot c + \alpha \cdot (t - a) + D(t)$, where α is the value of travel time, $t - a$ is the time spent in the bottleneck and $D(t) = \beta \cdot \max(0, t^* - t) + \gamma \cdot \max(0, t - t^*)$ is a convex function capturing the cost of being early or late relative to some preferred arrival time t^*. This model is often called the "$\alpha - \beta - \gamma$" model and the preferences "$\alpha - \beta - \gamma$" preferences. The Vickrey formulation of scheduling preferences is additively separable in trip duration and arrival time and it is linear in trip duration. So it is clear that the distance c to the bottleneck does not matter for the Vickrey analysis of how travellers time their arrival at the bottleneck and the ensuing congestion.[1]

It is not generally true that the distance from trip origins to the destination is irrelevant for the timing of trips. Consider a traveller who always prefers to depart later and always prefers to arrive earlier. Faced by a fixed trip duration that is independent of the departure time, such a traveller will optimally time his trip such that his marginal utility of being at the origin at the departure time equals his marginal utility of being at the destination at the arrival time. If the marginal utility at the origin is decreasing and his marginal utility of being at the destination is increasing, then an increase in trip duration will cause him to depart earlier and to arrive later. In this way the distance can matter for the timing of trips. This paper concerns travellers with such scheduling preferences.

Congestion can arise when there is a bottleneck and many individuals who want to pass the bottleneck at the same time. It is not a sufficient condition for congestion to arise that travellers have similar scheduling preferences. Trip origins must also be located with similar distances to the bottleneck. If trip origins are sufficiently dispersed, then congestion does not arise as there is no overlap in the times when travellers want to pass the bottleneck. Hence it is clear that the spatial distribution of travel demand is a fundamental determinant of urban congestion. This observation stands in contrast to the standard urban model, where congestion increases with population dispersion.

This paper is the first to allow for spatial heterogeneity in the bottleneck model in a meaningful way. In our model, heterogeneity is induced by the structure of the city. A number of earlier contributions have considered preference heterogeneity

[1]The analysis of the bottleneck model has been developed and extended in various directions by Arnott, de Palma and Lindsey in a series of papers; notably Arnott et al. (1993). These authors use the above $\alpha - \beta - \gamma$ preferences or a version where the function $D(t)$ has a more general form. They always maintain linearity and additive separability of travel time and are hence unable to analyse the consequences of distance for congestion.

in the context of the bottleneck model (e.g., Vickrey, 1973; Arnott et al., 1994; van den Berg and Verhoef, 2011). These papers work in the context of linear separable Vickrey (1969) scheduling preferences and heterogeneity is introduced by varying $\alpha - \beta - \gamma$, while maintaining the ratio β/γ fixed for reasons of analytical convenience. Generally speaking, this sort of heterogeneity can induce travellers to sort according to the degree of closeness to the center of the congestion peak; in a two group case, sorting has the form that one group occupies a central time interval while the other group occupies the early and late shoulders. In contrast, this paper finds that travelers sort according to their distance to the bottleneck; this occurs both under no tolling and under optimal tolling, and the result is derived under quite general assumptions concerning scheduling preferences.[2] Hendrickson and Kocur (1981), Smith (1984), Newell (1987), and Arnott et al. (1994) consider the case of travellers with scheduling preferences, such as $\alpha - \beta - \gamma$, that are additively separable in trip duration and arrival time and who differ in their preferred arrival time. In that case, travellers sort according to their preferred arrival time, which is similar to what we obtain here. Kuwahara (1990) extends this to a geometry consisting of two residential areas and a CBD with bottlenecks in between. Travellers within each group then still sort according to their preferred arrival time, but a strict sequence does not hold for the two groups together. The present case is more involved, as travellers have different distances to the CBD as well as strictly concave and non-separable scheduling preferences. We show that the optimal arrival time a_*, in the absence of congestion, is increasing in distance c, such that also here travellers sort according to their preferred arrival time.

Daganzo (2007) and Geroliminis and Daganzo (2008) show that several aspects of congestion in an urban area can be described as a form bottleneck congestion. A space average of traffic measurements show that the trip completion rate is a stable inverse u-shaped function of the number of vehicles present in the system. Cars that have not yet completed their trips remain in the urban area, such that it is possible to think of the system as a generalized sort of queue. See Geroliminis and Levinson (2009). The bottleneck model supposes a constant trip completion rate and a queueing system that maintains a first-in-first-out queue.

2.2 Findings

This paper has introduced spatial heterogeneity into the bottleneck model such that it can be used to represent a city with a central bottleneck.

Our analysis first shows that under laissez-faire, travellers sort according to their distance to the bottleneck such that those who are closest to the bottleneck reach the destination first. However, in general there is not a monotonous relation-

[2] Lindsey (2004) considers more general heterogeneity with a finite number of user classes.

ship between distance and departure time; it is not necessarily the case that those who are located further away will depart earlier.

The paper goes on to consider equilibrium under socially optimal tolling at the bottleneck. The toll can be taken to be zero for the first and last travellers and strictly positive for everybody else. The optimal toll exactly removes queueing. The sequence of arrivals at the destination is preserved from the laissez-faire equilibrium.[3] However, in contrast to the Vickrey analysis with homogenous travellers, arrivals at the destination occur earlier in social optimum than under laissez-faire. When the use of toll revenues does not affect the utility of travellers, then the toll just represents a loss for them. This is compensated to some extent by a gain in utility. Comparing social optimum to laissez-faire reveals that those who are located furthest away from the bottleneck will experience a net gain, while those who are located near the bottleneck will experience a net loss.

A number of new insights are generated from the model. Perhaps the most important insight is that travellers located near the bottleneck will tend to lose from optimal tolling, while those located far away will tend to gain, when the use of toll revenues is not accounted for. The paper also shows that a reason for the congested demand peaks to be uni-modal can be found in the properties of equilibrium in combination with our general specification of scheduling preferences.

2.3 Scientific perspectives

The spatial distribution of travellers is a source of heterogeneity in the model. It would be of interest to introduce other sources of heterogeneity into the model. One issue would be the robustness of the sorting property. Another kind of extension would be to introduce risk into the model, for example in the form of random capacity (Arnott et al., 1999) or random queue sorting (de Palma and Fosgerau, 2011).

Perhaps the most interesting extension would be to make the location of individuals endogenous as in the Mirrlees (1972) standard urban model. This would tie together congestion dynamics and urban economic models. For example Arnott (1998) combines a model of urban spatial structure with the $\alpha - \beta - \gamma$ bottleneck model; optimal tolling does not change transport costs for travellers so when the revenues are not returned, optimal tolling will have no effect on urban structure. As Arnott (1998) notes, this is in contrast to urban economic models with static congestion. However, the Arnott (1998) result is a consequence of space essen-

[3] Arnott et al. (1991) consider a variant of the standard bottleneck model in which drivers have to park, and parking spots are located at varying distances from the CBD. In the laissez-faire user equilibrium, drivers park in order of increasing distance from CBD. The optimal location-dependent parking fee reverses this pattern. This contrasts with the present setting where optimal tolling does not change the order of arrivals.

tially being assumed away in the specification of preferences as described in the paper's literature review.

3 Alternatives to pricing

The research that is summarized in this section has been published in Transportation Research Part B in a paper entitled "How a fast lane may replace a congestion toll" (Fosgerau, 2011). This paper considers a fast lane scheme as a means to regulate congestion in a regularly occurring demand peak.

3.1 Background

The fast lane scheme plays explicitly on the dynamics of congestion, which makes the Vickrey (1969) bottleneck model an appropriate framework. The elements of the basic bottleneck model are a description of the queueing technology in the bottleneck, a continuum of identical travellers with scheduling preferences who have to pass the bottleneck, and the concept of Nash equilibrium in arrival times at the bottleneck.

The fast lane scheme allocates the bottleneck capacity to different classes of travellers. The scheme is the following.

A set of travellers is assigned to a priority group. Not all travellers can be given priority. A more than proportional share of capacity is reserved for the priority group. When the reserved capacity is not used, it is available for the nonprioritized travellers.

This is similar to, e.g., the check-in in airports with separate queues and servers for economy and business class passengers. Whenever the business class server is idle, it may serve passengers from the economy class queue. Another example is the HOV or HOT lanes as found on US motorways. Yet another example is the flows at different motorway on-ramps that could be given different priority (Shen and Zhang (2010) consider such a scheme). Even though such a scheme is called a fast lane scheme in this paper, the definition encompasses many other policies that do not involve the allocation of road lanes for different classes of vehicles; it is more general than allocation of lanes.

The paper compares the fast lane scheme to tolling. Like Arnott et al. (1990), this paper considers a coarse toll, which is a constant toll that applies only during part of the peak.[4] Arnott et al. (1990) found that Nash equilibrium under a coarse toll comprises a point mass in the arrival schedule at the time when the toll is lifted. This is an undesirable feature of their model as such point masses are

[4] Arnott et al. (1990) applied also a base toll level. This paper considers the coarse toll only under inelastic demand and so the base toll level does not matter.

physically implausible. The problem is avoided in this paper by a reformulation of the queueing technology. In this paper, the congestion technology is such that travellers who choose not to pay the toll can queue at the same time as travellers who are paying the toll pass the bottleneck. This is also true of the examples of fast lanes given above. In this case, a point mass in the arrival schedule does not arise. The analysis below uses the reformulated queueing technology and repeats the Arnott et al. (1990) analysis of a coarse toll under this assumption.

3.2 Findings

The first main result of this paper is that the fast lane scheme is always Pareto improving when demand is not price sensitive. There are no restrictions on how large the group of prioritized travellers should be as long as it is fixed exogenously. Prioritized travellers experience a strict utility gain while the properties of Nash equilibrium imply that nonprioritized travellers do not lose. It is significant that this occurs even when travellers are homogenous and there are no toll payments. This robustness is very desirable since it means that a regulator needs little information to implement the scheme and be certain to achieve a welfare improvement. In fact, the regulator can monitor traffic in real time and assign capacity accordingly. This is consistent with the way the fast lane scheme is formulated in the model. With price sensitive demand, the fast lane scheme is still welfare improving if the price elasticity of demand is not too high and the share of prioritized travellers is not too large.

The second main result of this paper is that the fast lane scheme can reproduce the equilibrium arrival pattern of the optimal coarse toll when demand is not price sensitive.[5] In fact, the scheme can reproduce the equilibrium arrival pattern of any coarse toll, provided that the tolling interval is the same as the arrival interval that a prioritized group would endogenously select. This is significant since the fast lane scheme has a number of advantages over tolling. First, the fast lane scheme is always welfare improving and can be adjusted in real time. In order to set the right coarse toll it is necessary to know exactly when to start and when to end the tolling interval. Mistakes will reduce the welfare gain from tolling and can even lead to a welfare loss. Second, it is plausible that system costs can be a lot lower for a fast lane scheme than for a toll as the fast lane does not involve any payment. Finally, as there is no payment, a fast lane scheme may be more acceptable to travellers than tolling. Within the simple theoretical model

[5] Using $\alpha - \beta - \gamma$ scheduling preferences, Knockaert et al. (2010) show that the coarse charge user equilibrium can be obtained by barring a certain group from travelling during the charging period and letting the remainder of drivers travel in that period without paying any charge. The present fast lane scheme does not have to designate specific time periods for specific groups of travellers. Moreover, the result is shown to hold for quite general scheduling preferences.

presented here, prioritized travellers would be strictly better off under the fast lane scheme than under no policy, while the remaining travellers would be indifferent. In contrast, all travellers would be indifferent between tolling and no policy as toll revenues are not returned to travellers. This property of fast lanes may explain why fast lanes have been introduced while there is generally a reluctance to introduce tolls.

A notable feature of the present paper is the formulation of scheduling utility which generalizes those employed by Vickrey (1969, 1973), Arnott et al. (1990, 1993), and many others. Here, scheduling utility is taken to be a strictly concave function of times at which the trip starts and ends. Travellers prefer to depart later and to arrive earlier. For any fixed travel time there is a unique preferred departure time. These assumptions are sufficient for the results of this paper. The paper establishes that the socially optimal fast lane scheme achieves more than half the welfare gain of the socially optimal continuously time varying toll. This generalizes the parallel result by Arnott et al. (1990) for the coarse toll under their first-in-first-out congestion technology to the present formulation of scheduling utility combined with parallel queueing.

3.3 Scientific perspectives

It is straightforward but tedious to generalize the results of this paper to tolls with more steps and fast lane schemes with more user classes.[6] The general conclusion remains that fast lanes can achieve the same benefits as step tolls when demand is not price sensitive. It is also straightforward to see that a sequence of step tolls, and hence a sequence of fast lane schemes, can be constructed that approach the optimal time varying toll. In the limit, the step toll would become the optimal continuously varying toll while the fast lane scheme would become equivalent to allocating a specific time slot to every traveler.

A potentially useful feature of the fast lane scheme is its robustness. As long as demand is not too elastic, or as long as the share of prioritized travellers is no too large, then any fast lane scheme satisfying the conditions set up in the paper is welfare improving. If demand is not price sensitive, then any such fast lane scheme is Pareto improving. An interesting direction for further inquiry is how this robustness can be utilized. Is it the case that the fast lane scheme retains its favorable properties when some element of stochasticity is introduced into the model?

[6] Laih (1994) showed that it is straightforward to extend the coarse toll to a multistep toll. It is similarly straightforward to extend a fast lane scheme in this way. Laih (1994) did not recognize that it was necessary to reformulate the queueing technology in order to obtain his results. This was rectified in Laih (2004).

4 The nature of scheduling preferences

This research appears so far in a working paper "Endogenous scheduling preferences and congestion". It concerns a version of the model in which travellers do not have exogenous scheduling preferences. Instead they care about leisure and consumption. Both (effective) leisure and consumption are produced under increasing returns to scale. This, in combination with bottleneck congestion, leads to scheduling preferences arising endogenously in equilibrium. The importance for policy of this insight lies in the possibility that policies that affect congestion also affect scheduling preferences. This makes the effect of policies harder to assess ex ante.

4.1 Background

Static models of congestion use a static representation of demand in combination with a technology that relates travel cost to the number of travelers (see, e.g., Small and Verhoef, 2007). In essence, these models are based on the view that congestion occurs simply because many people want to travel to the same place, using the same infrastructure.

More recently, dynamic models of congestion take into account that congestion varies continuously over time and that conditions at one time affect congestion at later times. These models essentially view congestion as arising because many people want to go to the same place at the same specific time. For example, in the well-known Vickrey (1969) bottleneck model, this desire is hard-wired in individuals' utility functions through the concept of scheduling preferences for being at specific places at specific times; it is those preferences that explain the occurrence of congestion.

However, it seems plausible that the preferences for starting work at certain time are not innate, but arise for a reason. Thus Henderson (1981) posits agglomeration forces at the workplace to explain them, and Hall (1989) discusses rather generally how thick-market efficiencies lead to temporal agglomeration at various time scales. Such effects may explain the strong tendency for production to concentrate in the hours 9-12 a.m. and 1-5 p.m. We use this intuition to motivate an assumption that worker productivity increases in the number of people at work.

We apply the same reasoning to time off work, assuming that productivity in the production of effective leisure increases in the number of people off work at any given time. This is a reasonable assumption considering that many leisure activities are social, and others involve family members caring for each other.

In this paper we take workers simply to have preferences defined over leisure and consumption. We consider the morning commute with bottleneck congestion between home and work. Agglomeration economies at home and at work lead to

temporal agglomeration which in turn entails congestion. Thus, we present a view in which congestion occurs because people want to be at a place at the same time as other people are there.

4.2 Findings

It turns out that scheduling preferences of the kind assumed by Vickrey arise endogenously in equilibrium. That is, an individual taking equilibrium as given will appear to have scheduling preferences in the form of a utility function that depends on when the commute starts and ends. These scheduling preferences belong to a general class that, as far as we are aware, comprises all those specifications that have been considered by Vickrey and later authors in the context of the bottleneck model. We derive some properties of Nash equilibrium for this general class, in order to compare its predictions to those of our model, where the scheduling preferences arise endogenously. This allows us to evaluate the errors that result if policies aimed at regulating congestion are developed assuming incorrectly that scheduling preferences are exogenous.

We find that the assumption of exogenous scheduling preferences would lead an analyst to underestimate the benefit of congestion tolling. If the use of toll revenues does not affect workers, then an analyst relying on a Vickrey-like model would find workers to be indifferent between the situations with no or optimal tolling. This sort of conclusion is not available in the model with endogenous scheduling preferences, where travellers may either gain or lose from optimal compared to no tolling.

The results concerning optimal capacity provision and the marginal external cost of congestion are also ambiguous in general concerning the relative sizes of these under endogenous and exogenous scheduling preferences. We examine these and other properties of the two models using numerical simulations with a Cobb-Douglas utility function and simple power functions to describe agglomeration.

A few previous contributions analyze congestion and agglomeration economies in combination. Henderson (1981) analyses scheduling of work hours and work trips, based on the effect of agglomeration economies at work on wages and exogenous preferences concerning the timing of leisure. Wilson (1988) finds empirical evidence supporting the idea behind Henderson's model, namely that agglomeration economies at work cause workers to earn more if they start work during peak hours. Arnott (2007) reviews these papers and further applications, while adding his own innovation (still within a static framework) by allowing aggregate labor supplied to be affected by congestion tolls via a reduction in their net wage.

We employ a more general specification of agglomeration economies at work and introduce agglomeration economies in the production of leisure. Furthermore,

we substitute bottleneck queuing for Henderson's model of flow congestion; we thereby bypass certain inconsistencies that can arise between flow congestion in a dynamic setting and traffic dynamics (Chu, 1995), and take advantage of analytical advances that have accumulated in the many papers applying bottleneck queuing to analyze equilibrium scheduling (e.g., Vickrey, 1969, 1973; Newell, 1987; Fargier, 1983; Arnott et al., 1993).

This paper has presented a dynamic model of traffic congestion in which scheduling preferences arise endogenously. A naive analyst - observing equilibrium and assuming scheduling preferences to be exogenously given - would then make errors in predicting the effect of policies such as capacity expansion and tolling. The naive analyst would fail to identify one cost of queueing, namely the decrease in productivity of work and leisure that follows when some are stuck in traffic. Hence such an analyst would underestimate the benefit of a toll that removes queueing. Also, for some parameter sets, such an analyst would apply a toll schedule and/or aim for a departure pattern that is quite far removed from the optimal one. So a take-away of this paper for policy is that a gradual approach to introducing a policy such as road pricing is advisable, since that allows the consequences to be observed as one goes along.

The model with endogenous scheduling preferences generates an equilibrium that is indistinguishable from a model with exogenous scheduling preferences. It is hence not possible to falsify the latter model using only observation of individual choices in a single equilibrium; rather, in order to identify endogeneity, it is necessary to compare different equilibria. It may be possible to employ such an identification strategy empirically, for example by using capacity expansion or the introduction of a road pricing scheme as an exogenous instrument in an empirical investigation explaining variations in the temporal shape of the morning peak.

Humans are social animals and so it is entirely natural that the scheduling preferences of one individual should depend on the scheduling choices of others. We have shown that it is possible to model such a situation and that this interdependence affects transportation policy.

4.3 Scientific perspectives

There is a literature on social interactions (Manski, 2000), and traffic congestion may be viewed as an example of a social interaction. The social interactions in our model may be interpreted as occurring roughly at the level of a city. However, it seems likely that smaller scale interactions are also relevant. It would be interesting to develop, both theoretically and empirically, models of such interactions down to the scale of appointments between small groups of travelers.

5 Models for dynamic assignment and the nature of traffic equilibrium

An outstanding problem in transportation economics is the notion of traffic equilibrium. Consider the case where a number of individuals each morning choose departure time and route when going to work on a congested road network. This situation may be considered as a game, where each individual's utility (generalized cost, including time, monetary cost etc.) is affected by the strategies (departure time and route given information at hand) of the other individuals travelling in the morning commute.

A Nash equilibrium occurs when no commuter can be better off by unilaterally changing departure time and route, taking the strategies of all other commuters as granted. The Nash equilibrium is an extremely natural concept. It is also very useful since it allows us to predict the aggregate response to changes in policy. So it is very relevant to ask whether Nash equilibrium is a realistic description of actual behaviour. That requires that there is some process that can lead actual travellers to the equilibrium.

We can imagine that commuters are able to observe their own utility associated with the choice of departure time and route they made yesterday, under the conditions that prevailed yesterday. They may also be assumed to be able to observe the utility they would have acheived under alternative choices. Based on information of this kind, they make a new choice of departure time and route today. All commuters update their choices in this way and so the aggregate morning commute changes from day to day. The question is then whether such a process reaches a stable situation. Does simple heuristic learning rules exist that individual commuters can use to update their choice such that the aggregate converges towards an equilibrium? If not, then we (as researchers) have to reconsider our understanding of what it is that we observe and find new ways of making predictions.

Let us here briefly summarize what is known about the existence of convergent algorithms and learning equilibrium in this class of games. First, consider a static routing game in a network where players choose a route from origin to destination which is fixed (no en-route adaptation) and where all individuals travelling a particular link affect each other symmetrically. The timing of events is not important in this situation. This is a standard congestion game which has a number of important and nice features. For this class of games there exists a very intuitive and appealing learning rule. As individual simply update their path towards better paths, this process will converge towards a Nash equilibrium from any starting point. This follows from the fact that the game is a potential game, and the learning process will move towards a local optimum of the potential function. The potential function is a macro concept: it captures the state of the game

on a macroscopic level as individuals take decisions at the micro level. On a fundamental level, the potential function can be found by taking the limit of the dynamic process at the micro level. Mathematically, this limit between micro and macro is similar to the limit between quantum mechanics and classical mechanics in physics. In physics, it is known as the correspondence principle: we are able to analyze the behavior of the system on the macro scale (equilibrium) without have to worry about the details of the behavior on the micro scale (individual behavior). This is a very useful property when analyzing proposed policy measures (Karlstrom, 2012).

Second, consider the departure time game, where individuals only choose departure time. This game is not a potential game, but it still has some interesting properties. As has been shown in Hu and Fosgerau (2012), this game can be formulated as a stable game. This means that processes exist that converge to an equilibrium but, as far as we know, it is not the Nash equilibrium discussed above, but a noisy, probabilistic equilibrium.

Third, in a recent paper Young and Pradelski (2010) show that a particular heuristic simple learning mechanism will converge to a socially optimal Nash equilibrium in any game that exhibits at least one pure strategy Nash equilibrium. This shows that it may be possible to devise algorithms that calculate the social optimal Nash equilibrium in the static congestion game above, which does exhibit pure strategy Nash equilibria.

Finally, consider a dynamic routing game where the timing of entering links determines how different players asymmetrically affect the travel time of other players. For instance, one may assume that cars behind do not affect the travel time of cars further ahead.[7] This game is not a potential game, and much less is known about convergent algorithms or learning mechanisms that converge towards an equilibrium. In this dynamic routing game, it is unknown whether there exists a pure strategy equilibrium. It is also unknown whether it is a stable game. Likewise, less is known when introducing departure time into the dynamic routing game.

In summary, policy analysis so far has simply taken for granted that Nash equilibrium is a good description of what we observe in reality and has used that concept to predict the effect of policies. The strand of research discussed in this section investigates whether Nash equilibrium is the relevant equilibrium concept. The insights regarding convergence of learning mechanisms is also useful for devising algorithms that compute equilibrium in traffic simulation models.

[7] This is not universally true in all networks.

References

Arnott, R. A. (1998) Congestion Tolling and Urban Spatial Structure *Journal of regional science* **38**(3), 495–504.

Arnott, R. A. (2007) Congestion tolling with agglomeration externalities *Journal of Urban Economics* **62**(2), 187–203.

Arnott, R. A., de Palma, A. and Lindsey, R. (1990) Economics of a bottleneck *Journal of Urban Economics* **27**(1), 111–130.

Arnott, R. A., de Palma, A. and Lindsey, R. (1993) 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, A. and Lindsey, R. (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, A. and Lindsey, R. (1999) Information and time-of-usage decisions in the bottleneck model with stochastic capacity and demand *European Economic Review* **43**(3), 525–548.

Arnott, R., de Palma, A. and Lindsey, R. (1991) A temporal and spatial equilibrium analysis of commuter parking *Journal of Public Economics* **45**(3), 301–335.

Chu, X. (1995) Alternative congestion pricing schedules *Regional Science and Urban Economics* **29**(6), 697–722.

Daganzo, C. F. (2007) Urban gridlock: Macroscopic modeling and mitigation approaches *Transportation Research Part B: Methodological* **41**(1), 49–62.

de Palma, A. and Fosgerau, M. (2011) Random queues and risk averse users. DTU Transport.

Fargier, P. H. (1983) Effects of the choice of departure time on road traffic congestion University of Toronto Press Toronto.

Fosgerau, M. (2011) How a fast lane may replace a congestion toll *Transportation Research Part B* **45**(6), 845–851.

Fosgerau, M. and de Palma, A. (2012) Congestion in a city with a central bottleneck *Journal of Urban Economics* **71**(3), 269–277.

Geroliminis, N. and Daganzo, C. F. (2008) Existence of urban-scale macroscopic fundamental diagrams: Some experimental findings *Transportation Research Part B: Methodological* **42**(9), 759–770.

Geroliminis, N. and Levinson, D. M. (2009) Cordon pricing consistent with the physics of overcrowding *Proceedings of the 18th International Symposium on Transportation and Traffic theory* .

Hall, R. E. (1989) Temporal agglomeration *National Bureau of Economic Research* **3143**.

Henderson, J. V. (1981) The economics of staggered work hours *Journal of Urban Economics* **9**, 349–364.

Hendrickson, C. and Kocur, G. (1981) Schedule Delay and Departure Time Decisions in a Deterministic Model *Transportation Science* **15**(1), 62–77.

Hu, D. and Fosgerau, M. (2012) Departure time choice games *Technical report* DTU Transport, Denmark.

Karlstrom, A. (2012) Tropical economics - an essay on the correspondence principle in economics *Technical report* Royal Institute of Technology Stockholm, Sweden.

Knockaert, J., Verhoef, E. T. and Rouwendal, J. (2010) Bottleneck Congestion: Differentiating the Coarse Charge *Working Paper* .

Kuwahara, M. (1990) Equilibrium Queueing Patterns at a Two-Tandem Bottleneck during the Morning Peak *Transportation Science* **24**(3), 217–229.

Laih, C.-H. (1994) Queueing at a bottleneck with single- and multi-step tolls *Transportation Research Part A* **28**(3), 197–208.

Laih, C. H. (2004) Effects of the optimal step toll scheme on equilibrium commuter behaviour *Applied Economics* **36**(1), 59–81.

Lindsey, R. (2004) Existence, Uniqueness, and Trip Cost Function Properties of User Equilibrium in the Bottleneck Model with Multiple User Classes *Transportation Science* **38**(3), 293–314.

Manski, C. F. (2000) Economic Analysis of Social Interactions *The Journal of Economic Perspectives* **14**(3), 115–136.

Mirrlees, J. A. (1972) The Optimum Town *Swedish Journal of Economics* **74**(1), 114–135.

Newell, G. F. (1987) The Morning Commute for Nonidentical Travelers *Transportation Science* **21**(2), 74–88.

Shen, W. and Zhang, H. M. (2010) Pareto-improving ramp metering strategies for reducing congestion in the morning commute *Transportation Research Part A* **44**(9), 676–696.

Small, K. A. and Verhoef, E. T. (2007) *Urban transportation economics* Routledge London and New York.

Smith, M. J. (1984) The Existence of a Time-Dependent Equilibrium Distribution of Arrivals at a Single Bottleneck *Transportation Science* **18**(4), 385–394.

van den Berg, V. and 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**(7-8), 983–992.

Vickrey, W. S. (1969) Congestion theory and transport investment *American Economic Review* **59**(2), 251–261.

Vickrey, W. S. (1973) Pricing, metering, and efficiently using urban transportation facilities *Highway Research Record* **476**, 36–48.

Wilson, P. W. (1988) Wage variation resulting from staggered work hours *Journal of Urban Economics* **24**(1), 9–26.

Young, H. P. and Pradelski, B. S. (2010) Learning Efficient Nash Equilibria in Distributed Systems *Technical report* University of Oxford.

IV. TARIFICATION DU STATIONNEMENT COMME SUBSTITUT A LA TARIFICATION ROUTIERE DANS LE CADRE D'UN MODELE DYNAMIQUE AVEC CONGESTION

Mogens Fosgerau, Technical University of Denmark et Center for Transport Studies, Suède

André de Palma, Ecole Normale Supérieure de Cachan et Centre d'économie de la Sorbonne. CES, France

RESUME

On considère le modèle du goulot d'étranglement pour le pic du matin. Une tarification qui dépend continument de l'heure de la journée permet d'éliminer la congestion (et permet de réduire le coût généralisé des utilisateurs de moitié). Le coût supporté par l'usager est la somme du temps de trajets et du coût de déshorage associés aux arrivées précoces ou tardives à la destination.

On considère dans un premier temps le cas de la demande inélastique. La tarification optimale augmente d'abord au cours du temps et ensuite diminue. Elle est maximale pour l'automobiliste qui arrive à la destination juste à l'heure. On envisage dans cet article d'autres modes de tarification, dans lesquelles les usagers paient à la destination : en d'autres termes on étudie l'efficacité de la tarification des stationnements. On se place dans un cadre dynamique, de sorte que la tarification du stationnement dépend de l'heure d'arrivée des usagers à la destination. Ce type de tarification introduit évidemment des contraintes. Par exemple si un automobiliste A arrive plus tard que l'automobiliste B, le premier paiera davantage (ou la même chose) que le premier, si les temps de départ de la destination sont les mêmes.

On montre que la tarification du stationnement est nulle jusqu'à ce que la file d'attente commence et ensuite elle diminue avec le temps d'arrivée. On calcule la perte d'efficacité vis-à-vis de la solution optimale de premier rang (tarification de la route). Enfin, on analyse les pertes d'efficacité, si certain usagers ne peuvent être tarifés.

Lorsque la demande totale est élastique, la tarification en début de journée n'est plus nulle, mais fonction de l'élasticité de la demande au coût généralisé.

Mots clé: Tarification du stationnement, Congestion, Politique de tarification, Modèle dynamique.

Codes JEL: D00; D80

The dynamics of urban traffic congestion and the price of parking*

Mogens Fosgerau
mf@transport.dtu.dk

André de Palma
andre.depalma@ens-cachan.fr

September 19, 2012

Abstract

We consider commuting in a congested urban area. While an efficient time-varying toll may eliminate queuing, a toll may not be politically feasible. We study the benefit of a substitute: a parking fee at the workplace. An optimal time-varying parking fee is charged at zero rate when there is queuing and eliminates queuing when the rate is non-zero. Within certain limits, inability to charge some drivers for parking does not reduce the potential welfare gain. Drivers who cannot be charged travel when there is queuing. In some cases, interaction between morning and evening commutes can be exploited to remove queueing completely.

Keywords: parking; dynamic; congestion; urban; traffic
JEL codes: D0; R4

*Fosgerau: Technical University of Denmark; Centre for Transport Studies, Sweden; Ecole Normale Supérieure de Cachan, Centre d'économie de la Sorbonne, France. de Palma: Ecole Normale Supérieure de Cachan, Centre d'économie de la Sorbonne, France. We are grateful for comments from Robin Lindsey, Olivier Beaude, Jos van Ommeren, Dereje Fentie Abegaz and seminar participants at the 2011 Kuhmo-Nectar Conference and at the Tinbergen Institute. We are especially grateful for the very constructive comments that we received from the referees and the editor Dennis Epple, which have lead to considerable improvement and broadening of the scope of the paper. This research has been carried out under the following projects: Tarification des transports individuels et collectifs à Paris. Dynamique de l'acceptabilité : PREDIT and Ademe. Surprice project, Scheduling, trip timing and scheduling preferences, PREDIT.

1 Introduction

Traffic congestion is an economically important problem affecting cities everywhere. An average American household travels annually about 20,000 miles on roads and spends about 15% of income on road transportation.[1] In 2010, congestion in the US caused around 4.8 billion hours of travel delay and 1.9 billion gallons of extra fuel consumption with a total cost of $101 billion (Schrank et al., 2011). Thus, policies to reduce the cost of mobility by car are of first order importance. This paper considers the possibility of using parking fees rather than congestion pricing to regulate urban congestion by influencing the timing of trips.

Economists have advocated marginal cost pricing of road capacity as a means to improve efficiency for more than 100 years. However, very few cities have actually implemented congestion tolls, notably Stockholm, Singapore, and London. Congestion tolls have been proposed and then scrapped in many places, including New York, Hong Kong and Copenhagen. So there seems to be important political obstacles to congestion tolls and it is therefore of interest to look for alternative policies that can address road congestion.[2] It is natural to look at parking pricing, since parking is already priced almost everywhere. Another reason, noted by Shoup (2005), is that the technology needed to charge for parking is much simpler than that needed to charge for driving in congested traffic.

It is straightforward that the demand for trips to a city center is affected by the full price of the trip, including the price of parking. But the problem is not just the volume of traffic: the timing of demand is extremely important as is evident from the sharp demand peaks that characterize urban traffic. The physics of congestion

[1] http://nhts.ornl.gov/2009/pub/profile_2012.pdf and http://www.bls.gov/cex.

[2] De Borger and Proost (2010) discuss the political economy of road pricing.

implies that the amount of congestion delay is strongly dependent on the timing of trips. If only departures from home in the morning became more dispersed in time, then congestion delay could be much smaller while arrival times could be quite unaffected. So there is a large potential efficiency gain in the retiming of trips, even if the total traffic volume is unaffected. Congestion tolling aims to achieve such temporal dispersion by applying a toll that varies over time with the amount of congestion. The purpose of this paper is to explore the potential for time-varying parking pricing to achieve the same effect.

We use a generalized version of the Vickrey (1969) bottleneck model for this purpose (de Palma and Fosgerau, 2011a). The bottleneck model captures the essence of congestion dynamics, describing a continuum of drivers equipped with preferences regarding the timing of a trip to a common destination. This destination is located behind a bottleneck with a fixed capacity. If the rate at which drivers want to pass the bottleneck exceeds its capacity then delay results.[3] The delay is a pure loss and it could be reduced with no effect on arrival times if people could be induced to choose different departure times. A time-varying toll aims to induce such rescheduling. As long as it induces appropriate rescheduling of trips, it makes no difference where the toll is collected, it can be on any point of the trip.

In this paper we exploit that drivers park at the destination and pay a parking fee. We will mainly consider a parking fee that accumulates at a non-negative time-varying rate. This restriction fundamentally distinguishes such parking fees from congestion tolls. Congestion tolls may vary freely up and down and may be lower on the shoulders of the peak and high in the middle. A parking fee charged

[3]The bottleneck congestion technology is a means to represent city-wide congestion affecting all traffic and the bottleneck does not necessarily correspond to any single place in a city (Daganzo, 2007; Geroliminis and Daganzo, 2008).

at a positive rate during parking is always lower for later arrival times. As drivers differ in the time at which they pass the bottleneck, they differ also in the parking fee they pay. Therefore a parking fee can be used to induce rescheduling of trips but in a less flexible way than a toll.

In summary, parking fees seem to be much easier to introduce than congestion tolls. Like congestion tolls, parking fees may be used to disperse demand over time in order to reduce congestion and gain efficiency, but the efficiency gain may be limited by the restrictions inherent in typical parking fees. The objective of this paper is to present an analysis of parking fees as a means to affect the timing of road use and as an alternative to congestion tolls.[4]

We initially make assumptions that allow us to ignore the influence of the time of unparking. We may think of the destination as the workplace, such that the model describes the morning commute. For the morning commute, we find that the imposition of a parking fee causes the departure interval to occur later than it would in the absence of policy. This shift compensates the early drivers who pay more for parking than later drivers. The optimal parking fee implements a situation where every morning there is first an interval with a demand peak that involves queueing just like an unregulated equilibrium except that it does not involve everybody traveling in the morning. The optimal parking fee rate is zero during this interval such that the total parking fee is the same for all these drivers. The optimal parking fee becomes positive at the time when the queue has dissolved and is set such that zero queue is maintained during the remainder of the morning.

[4]The US Federal Highway Administration has a series of parking pricing projects under their value pricing pilot program (in San Francisco, San Diego, and New York) that include time-varying parking fee rates.

It is a recurring theme in the debate about charging for parking that some drivers cannot be charged since they have private parking available. In the current situation, it turns out they make no difference provided they can fit within the period where the optimal parking fee rate is zero and queueing occurs. Thus, within this limit, the existence of private parking does not affect the welfare gains that can be achieved from a parking fee.

Another way that drivers may escape the time-varying parking fee rate is through early bird specials, providing all day parking at a discounted price for drivers who arrive at a parking lot by a certain time such as 8 am. The paper characterizes the welfare maximizing combination of an early bird special with a time-varying parking fee rate.

After examining the morning peak, we show that the conclusions of the paper extend with few modifications to the evening commute, where parking is charged at the origin of the trip instead of at the destination.[5] The optimal parking fee affects the evening commute similarly to the morning commute, except that the order of the congested and uncongested intervals is reversed and the departure interval occurs earlier than it would in the absence of the parking fee.

The analysis so far ignores any interaction between the two commutes. The paper also analyzes a whole day with explicit interaction between the two commutes. Nonseparability between the morning and evening commutes implies that the morning commute can be affected via the evening parking fee and vice versa.

[5] de Palma and Lindsey (2002) compare the morning and the evening commute, assuming that scheduling utility is additively separable in travel time and delay, where delay is defined in terms of arrival time for the morning commute and in terms of departure time for the evening commute. Here, we apply a general form of scheduling preferences that applies to both the morning and the evening commute. The difference in principle between the two commutes is whether the parking fee is charged at the origin or at the destination of the trip.

It turns out the limitations involved in parking pricing as compared to freely time-varying congestion tolling can then be overcome, and in our stylized setting a parking fee scheme can be designed to remove congestion completely during both commutes simultaneously. This finding strengthens the case for using parking pricing to tackle urban road congestion.

The first to discuss regulation of parking in an economic context might be Vickrey (1954), who suggested time-varying parking fees as a means of regulating the use of parking space. Glazer and Niskanen (1992) present an analysis where parking fees are analyzed as a substitute for road pricing. They note that the idea rests on the assumption that an increase in the price of parking is equivalent to an increase in the price of a trip. However, this equivalence fails for people who can vary the length of time they park. Increasing the parking fee rate may induce drivers to park for a shorter time, thereby allowing more people to use parking spaces each day and thereby increasing traffic. However, Glazer and Niskanen (1992) do not consider congestion dynamics (see also Verhoef et al., 1995).

In a static simulation model, Calthrop et al. (2000) analyze the efficiency gains from parking fees and road pricing (a cordon toll). They find that these two policies are sub-additive: as roads are more efficiently priced, there is less need for pricing of parking. In contrast to us, they also find that second-best pricing of parking produces a higher welfare gain than a cordon charge around the simulated city. The explanation for this difference is that they consider the supply of parking but no congestion dynamics, where we take the supply of parking as given and consider how to exploit congestion dynamics using a time-varying parking fee.

Like us, Arnott et al. (1991) use the bottleneck model, but they consider a case where parking spaces are located between the bottleneck and the CBD, on

a line away from the CBD and where the parking cost varies according to the distance to the CBD. In their analysis, the parking fee does not depend on the length of time the vehicle is parked. Arnott et al. (1991) find that optimal location-dependent parking fees do not eliminate queueing, but induce drivers to park in order of decreasing distance from the CBD, thereby concentrating arrival times closer to work start times. They find that for most reasonable parameter values, the optimal location-dependent parking fee is at least as efficient as the optimal time-varying road toll. In contrast, in the present setting where parking is located at the destination and with temporal but not spatial variation in the parking fee, only a smaller share of the efficiency gain from the optimal road toll can be realized by a parking fee. Qian et al. (2012) present an analysis similar to Arnott et al. (1991) but with parking capacity provided in two parking lots, where the capacity and parking fee may be regulated.

Arnott and Rowse (2009) focus on different aspects of parking. They analyze parking in a spatially homogeneous downtown area. Drivers choose between curb-side and garage parking, and curbside parking is cheapest. Cruising for parking contributes to congestion and works to increase the full price for curbside parking until it equals the price of garage parking. Then increasing the curbside parking fee may generate an efficiency gain through reduction of cruising and the ensuing congestion and the efficiency gain may be large relative to the parking fee revenue. Other papers related to cruising include Douglas (1975), Arnott and Rowse (1999), Anderson and de Palma (2004), Arnott and Inci (2006), and Anderson and de Palma (2007). Van Ommeren et al. (2011) estimates the cost of cruising for the residents of Amsterdam. See also Proost and Van Dender (2008) and De Borger and Wuyts (2009).

Zhang et al. (2005) link the morning and evening commutes by treating the length of the work day as a decision variable in a model similar to ours. They do not analyze time-varying parking fees.

Section 2 introduces the model, Section 3 reviews the benchmark case of no policy, while Section 4 reviews the optimal time-varying toll at the bottleneck. Section 5 describes equilibrium under a parking fee, Section 6 considers the optimal parking fee, and Section 7 presents an example under specific assumptions about scheduling preferences. Section 8 considers the case when some drivers cannot be charged for parking and Section 9 characterizes social optimum including an early bird special. Section 10 discusses the evening commute while Section 11 considers the two commutes in combination. Section 12 concludes. Most proofs are relegated to the Appendix.

2 Model formulation

There is a continuum of mass $N > 0$ of drivers who all have to pass a congested bottleneck. They have identical preferences concerning the timing and cost of their trip expressed by the twice differentiable money metric utility $u(t, a) - \tau$, defined for all $t \leq a$ and τ, where t is the arrival time at the bottleneck, a is the exit time from the bottleneck and τ is the (monetary) cost of the trip. We speak of the length of the duration from t to a as the travel time or the bottleneck delay. We consider only costs in the form of a toll at the bottleneck or a parking fee at the destination. We refer to u as the scheduling utility.[6] Without loss of generality, t represents also the departure time and a the arrival time at the destination. It is also

[6] A simple version of scheduling preferences have the so-called $\alpha - \beta - \gamma$ form formulated by Vickrey (1969), estimated by Small (1982), and used by numerous authors since.

useful to define the schedule delay utility $v(t) \equiv u(t,t)$, which is the scheduling utility that is obtained when travel time is zero. Throughout this paper we make the following assumptions regarding the scheduling utility.

Assumption 1 *Marginal scheduling utility satisfies* $u_1 > 0$ *and* $u_2 < 0$. *Schedule delay utility* $v(t)$ *is strictly quasiconcave and attains maximum* $v(t_*)$ *at* t_*.

The assumption first requires that drivers always strictly prefer to depart later and to arrive earlier, no matter when they depart and arrive. The assumptions on v will ensure the uniqueness of equilibrium in the model.

The bottleneck has a capacity of ψ cars per time unit. Cars who have not yet been served wait before the bottleneck, which serves travelers in the sequence of arrival (first-in-first-out). The bottleneck capacity is always used if there are cars waiting before it. The physical extension of the queue has no consequences, we say the queue is vertical.

Cumulative departures are denoted $R(\cdot)$ and departures take place during an interval $[a_0, a_1]$. When $R(\cdot)$ is differentiable, we let $\rho(\cdot) = R'(\cdot)$ be the departure rate. If queueing begins at time a_0 and there is still queue at time t, then the queue length at time t is $R(t) - \psi(t - a_0)$ and the driver departing from home at time t exits the bottleneck at time

$$t + \frac{R(t) - \psi(t - a_0)}{\psi} = \frac{R(t)}{\psi} + a_0. \tag{1}$$

After passing the bottleneck, cars enter a parking space, which is vertical like the queue. Drivers pay a parking fee at a positive time-varying rate from the time of arrival at the parking lot until a time Ω which is the same for all drivers. The

latter assumption allows us to focus attention on the interaction of the parking fee with the departure time and rules out any interaction with the later departure from the parking space. Specifically, it does not require that all cars have to leave the parking space at time Ω. It is sufficient if utility is a separable part of a more comprehensive utility that also describes preferences regarding times later than Ω, Later, in Section 11, we shall consider a case without separability between the two commutes.

We will not consider situations involving mass departures and so the cumulative departure rate will be invertible. For this reason and since the queue is first-in-first-out, we can make a change of variable and equivalently define the parking fee rate $\pi(\cdot) \geq 0$ in terms of the departure time t. The parking fee for a driver departing and arriving at the bottleneck at time t is then $P(t) = \int_t^\Omega \pi(s)\, ds$ and we consider only π such that $P'(t) = -\pi(t)$.[7]

The analysis considers Nash equilibrium, which is defined by the property that, given the departure schedule R, no driver is able to strictly increase utility by unilaterally changing departure time. All drivers achieve the same utility in Nash equilibrium. The welfare measure employed is the equilibrium utility of drivers times the number of drivers plus the revenue from any toll or parking fee. Since utility is the scheduling utility minus the monetary cost, the welfare measure is equal to the total scheduling utility obtained by drivers.

[7]This avoids having to deal with issues related to sets of measure zero.

3 No policy equilibrium

Consider as an introduction the case of no policy. Nash equilibrium has arrivals at the bottleneck during an interval $[a_0, a_1]$, the endpoints of this interval are endogenous and determined in equilibrium. Equilibrium requires that there cannot be unused capacity during this interval, that there cannot be queue at the time of the last departure and that the utility of the first and last drivers to depart are the same.[8] Then the departure interval is uniquely determined by the conditions

$$v(a_0) = v(a_1),$$
$$a_1 = a_0 + N/\psi.$$

The conditions imply that $a_0 < t_* < a_1$, since v is strictly quasiconcave. There is always queue in the interior of $[a_0, a_1]$. The equilibrium is illustrated in Figure 1.

In equilibrium, the number of departures $R(t)$ that have occurred at time t can be determined using (1) by the equation

$$v(a_0) = u\left(t, \frac{R(t)}{\psi} + a_0\right). \tag{2}$$

This determines $R(t)$ since $a \to u(t, a)$ is invertible for all t. Moreover, differentiating (2), the departure rate is given by

$$\rho(t) = -\psi \frac{u_1\left(t, \frac{R(t)}{\psi} + a_0\right)}{u_2\left(t, \frac{R(t)}{\psi} + a_0\right)} > 0.$$

[8] If there were unused capacity with departures before or after then some driver could move into the gap and gain. If there were queue at the time of the last departure, then the last driver could postpone departure without affecting arrival which would yield a gain.

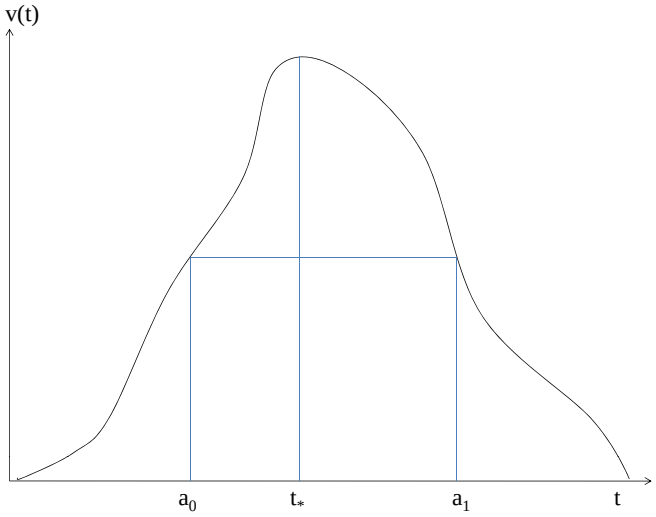

Figure 1: Schedule delay utility and no policy equilibrium

Here and later, the departure rate is determined almost everywhere.

Figure 2 shows the cumulative departures R as well as the number of cars that have passed the bottleneck $\psi(t-a_0)$. The vertical distance between the two curves corresponds to the length of the queue and the horizontal distance corresponds to the delay in the queue.

4 The optimal time-varying toll at the bottleneck

It is well known that a time varying toll can achieve maximum efficiency by removing the incentive to queue (Vickrey, 1969, 1973; Arnott et al., 1993; de Palma and Fosgerau, 2011b). The efficient toll is charged at the bottleneck at the time varying rate $\tau(t)$. Since total demand is assumed to be completely inelastic, we can set $\tau(a_0) = 0$ at no loss of generality. Efficiency requires $v(a_0) = v(a_1)$ so the efficient toll leaves the departure interval $[a_0, a_1]$ unchanged relative to the no

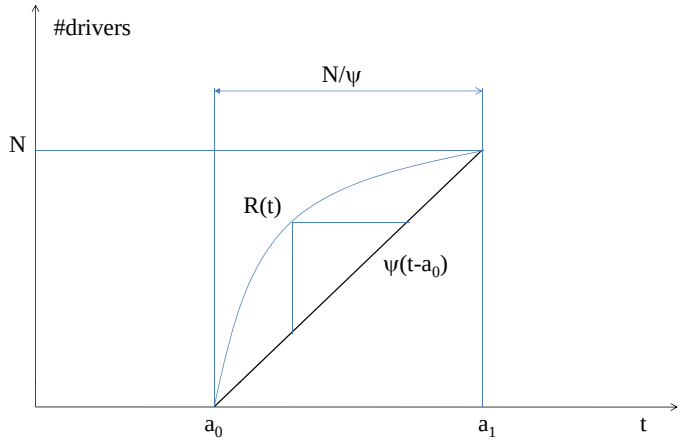

Figure 2: No policy equilibrium

policy equilibrium while maintaining the departure rate at $\rho(t) = \psi$. This requires $\tau(t) = v(t) - v(a_0)$. It follows that the efficient toll inherits strict quasiconcavity from v. Moreover, $a_0 < t_* < a_1$, and the efficient toll is increasing on $[a_0, t_*]$ and decreasing on $[t_*, a_1]$. The revenue from the efficient toll is

$$TR = \psi \int_{a_0}^{a_1} (v(t) - v(a_0))\, dt.$$

Drivers achieve the same utility in equilibrium as under no policy and hence the revenue from the efficient toll is equal to the welfare gain.

5 Parking fee equilibrium

Consider now a parking fee $P(t) = \int_t^{\Omega} \pi(s)\, ds$, where Ω is larger than any departure time. By definition it is decreasing as a function of arrival time (since

$P'(t) = -\pi(t)$) and hence it cannot replicate the efficient toll, which is increasing early in the peak.

Some basic properties of equilibrium are given in the following theorem. The proof is included here in the main text since it is helpful in motivating the conditions of the theorem.

Theorem 1 *Consider a parking fee schedule $P(\cdot)$ with*

$$b_0 < t_* < b_1 \tag{3}$$

$$v(t) - P(t) \geq v(b_0) - P(b_0) \Leftrightarrow t \in [b_0, b_1] \tag{4}$$

$$b_1 = b_0 + \frac{N}{\psi} \tag{5}$$

$$\pi(t) + u_2(t,t) < 0. \tag{6}$$

Then $\Delta \equiv P(b_0) - P(b_1) \leq \Delta^* \equiv v(t_*) - v(a_0)$ and there exists a unique departure time equilibrium solution defined on $[b_0, b_1]$. b_0 increases strictly as a function of Δ as Δ ranges over $[0, \Delta^*]$.

Proof. That $\Delta \leq \Delta^*$ follows from (3) and the quasiconcavity of v. Condition (4) ensures that nobody will want to depart outside $[b_0, b_1]$ and condition (5) ensures that all cars fit within this interval with capacity utilized throughout. Existence and uniqueness of equilibrium then follows if there exists a unique departure rate maintaining constant utility for departures in $[b_0, b_1]$. Condition (4) ensures that utility can be constant in equilibrium for departures within $[b_0, b_1]$ with non-negative queue length and this ensures that capacity is fully utilized during $[b_0, b_1]$. The equilibrium queue length exists uniquely and then so does the equilibrium departure rate from home. Condition (6) ensures that the equilibrium departure rate

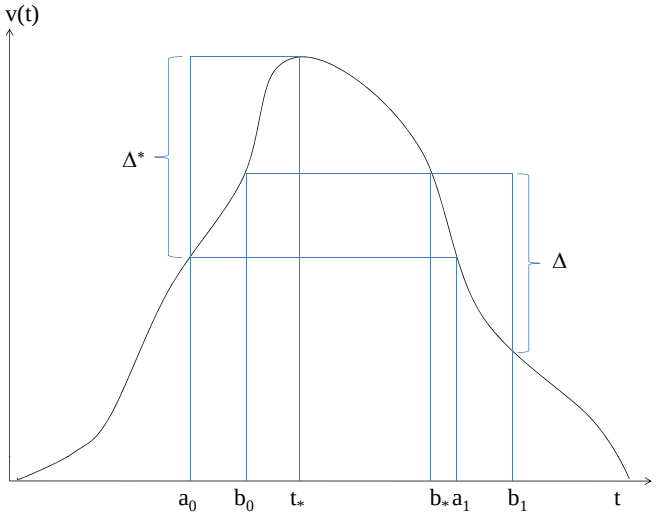

Figure 3: Equilibrium with parking fee

from home is strictly positive. The final conclusion of the theorem follows from the strict quasiconcavity of v. ■

Define for convenience b_* as the unique time $b_* > t_*$ where $v(b_0) = v(b_*)$. The equilibrium is illustrated in Figure 3.

6 Optimal parking fee

Fixing the difference Δ at some value and finding the corresponding departure interval $[b_0, b_1]$, welfare is maximized for a parking fee that extracts maximal revenue while satisfying the condition (4).

Find the unique $b_* > t_*$ with $v(b_0) = v(b_*)$ (see Figure 3). Let $P(t) = P(b_0)$ for $t \in [b_0, b_*]$. This satisfies the conditions of Theorem 1. It is also true that $R(b_*) = \psi(b_* - b_0)$, such that the queue is exactly gone at time b_*.

During the remaining time $[b_*, b_1]$ let $P(t) = v(t) - v(b_0) + P(b_0)$. This also

satisfies the conditions of Theorem 1.

With this fee, utility is constant during $[b_*, b_1]$ so there can be no queue. Therefore it is not possible to extract further revenue during this interval. We have therefore established the optimal parking fee conditional on a value of Δ.

Assume without loss of generality that $P(b_1) = 0$. The welfare function defined in terms of Δ is

$$W(\Delta) = \psi(b_* - b_0)\, v(b_0) + \psi \int_{b_*}^{b_1} v(t)\, dt. \tag{7}$$

We can find the optimal value of Δ as given in the following theorem. All proofs of this and theorems following below are given in the Appendix.

Theorem 2 *The optimal parking fee rate is*

$$\pi(t) = \begin{cases} 0, & t \in [b_0, b_*], \\ -v'(t), & t \in]b_*, b_1]. \end{cases}$$

Assume further that $v(\cdot)$ is concave. Then the welfare function $W(\cdot)$ is quasiconcave on $]0, \Delta^*[$, the welfare maximizing value of Δ exists, is unique and satisfies $\Delta = (b_* - b_0)\, v'(b_0) \in]0, \Delta^*[$.

The first statement of this theorem is that the optimal parking fee rate is zero during the interval $[b_0, b_*]$, which is the interval where there is queue under the optimal parking fee. Thus all drivers in this interval pay the same total amount for parking. The parking fee is concentrated on the interval $]b_*, b_1]$, where it ensures that there is no queue.

Figure 4 illustrates the evolution of queue lenght under no policy and under

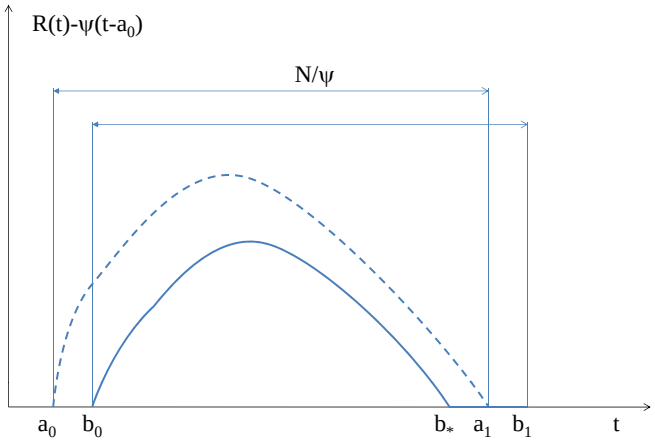

Figure 4: Evolution of the queue under no policy and under the optimal parking fee

the optimal parking fee. The dashed line shows that under no policy the queue first builds and then dissipates between times a_0 and a_1 and that these times span a duration of N/ψ time units. Queueing begins later at time b_0 under the optimal parking fee and it also ends earlier at time b_*. Departures continue during $[b_*, b_1]$ at the capacity rate such that there is no queue during this interval. The latest arrival at time b_1 occurs later than it would under no policy.

7 Linear specification

This section specializes results to the case of so-called $\alpha - \beta - \gamma$ preferences (Vickrey, 1969; Arnott et al., 1993). Let $v(a) = \beta \cdot \min(a, 0) - \gamma \cdot \max(a, 0)$ and let utility be $u(t, a) - \tau = v(a) - \alpha \cdot (a - t) - \tau$. Then α is the value of time, the marginal cost of lateness is γ and the marginal cost of earliness is β.

Let $0 < \beta < \alpha, 0 < \gamma$, as is typically assumed (Small, 1982). Then u satisfies the requirements stated in Section 2. The following proposition, proved in the Appendix, provides the optimal welfare gain in terms of the welfare function W, defined in (7). It thus states that the optimal welfare gain is obtained when the difference Δ in parking fee for the first and last drivers is equal to $\frac{\beta\gamma}{\beta+\gamma}\frac{N}{\psi}$. The proposition also evaluates the welfare gain in that case.

Proposition 1 *The optimal parking fee leads to a welfare gain of*

$$W\left(\frac{\beta\gamma}{\beta+\gamma}\frac{N}{\psi}\right) - W(0) = \frac{N^2}{\psi}\frac{\beta^2\gamma}{2(\beta+\gamma)^2}.$$

The interval without queuing has duration

$$b_1 - b_* = \frac{\beta}{\beta+\gamma}\frac{N}{\psi}.$$

Thus, a share $\frac{\beta}{\beta+\gamma}$ of drivers arrive during the later period when the parking fee removes queueing. The maximal welfare gain corresponds to a share of $\frac{\beta}{\beta+\gamma}$ of the maximal welfare gain that can be obtained by a time-varying toll at the bottleneck and the share is strictly less than $1/2$ when $\beta < \gamma$ as would commonly be assumed. It is also straightforward to verify that the revenue from the optimal parking fee corresponds to the same share of $\frac{\beta}{\beta+\gamma}$ of the revenue from the optimal time varying toll. The optimal coarse toll, i.e. a toll that has only two values, captures half the welfare gain that can be obtained by the optimal time-varying toll (Fosgerau, 2011) and so the optimal parking fee approaches this welfare gain when β is close to γ. These results are invariant under proportional changes in (β, α, γ). A value of γ/β in the range $2-4$ is reasonable and leads to an optimal

welfare gain in the range $[0.08, 0.11] \cdot N^2/\psi$, and this is between one fifth and one third of the gain that could be obtained by the optimal time varying toll or between two fifths and two thirds of the gain that could be obtained by the optimal coarse toll.

8 Private parking

We consider now a situation where some drivers cannot be charged for parking. This could be because they have private parking available that cannot be charged by the public authority. Let $N = N_c + N_u$, where N_c is the number of drivers that can be charged and N_u is the number of drivers that cannot be charged. Drivers are otherwise identical and they cannot affect whether they can be charged for parking or not. This assumption enables us to focus on the direct effects of parking fees without having to worry about selection into groups. Charged and uncharged drivers share the same queue at the bottleneck.

Let the departures of uncharged drivers take place during S^u with $Conv\left(S^u\right) = [b_0^u, b_1^u]$ and similarly let departures for charged drivers take place during S^c with $Conv\left(S^c\right) = [b_0^c, b_1^c]$.[9] Let $b_0 = \min\left(b_0^u, b_0^c\right)$, and $b_1 = \max\left(b_1^u, b_1^c\right)$. The following theorem establishes some properties of Nash equilibrium.

Theorem 3 *Consider a parking fee satisfying the assumptions (3-6) of Theorem 1. Then, in Nash equilibrium, capacity is fully utilized during* $[b_0, b_1]$ *and* $b_1 = b_0 + N/\psi$. *Uncharged drivers depart within the interval* $[b_0, b_*]$ *with* $b_* < b_1$.

The theorem shows that uncharged drivers depart within the period when there

[9] $Conv\left(\cdot\right)$ denotes the convex hull; the convex hull of a set on the real line is an interval.

is congestion and schedule delay utility v is largest. Some of the charged drivers are induced to travel later and they all achieve lower utility.

Let N_u and N_c be given. We may then ask what is the optimal charge. Using Theorem 2 and the preceding discussion, the optimal charge that charges N_c drivers satisfies $v\left(b_0^u\right) = v\left(b_0^u + \frac{N_u}{\psi}\right)$ and $\pi(t) = -v'(t)$ for $t > b_0^u + \frac{N_u}{\psi} \equiv b_1^u \equiv b_0^c$. Departures of charged drivers take place from b_0^c to $b_1^c = b_0^c + N_c/\psi$. We have $\Delta = P\left(b_1^c\right) - P\left(b_0^c\right) = v\left(b_1^c\right) - v\left(b_0^c\right)$. In case Δ is larger than its optimal value from Theorem 2, then there can be an early period with zero charge for charged drivers such that the optimum outcome is obtained. If on the other hand, the number of drivers that can be charged is less than the optimal number, then the optimal charge under this restriction is the one just described.

9 Early bird specials

Early bird specials are common in cities around the world (Victoria Transport Policy Institute, 2012) and they are targeted at commuters. Early bird specials provide all day parking at a discounted price for all-day parkers who arrive at a parking lot by a certain time such as 8 am. This section presents an analysis of how early bird specials can be used to reduce traffic congestion and improve welfare. An early bird special is given by (N_{eb}, a_{eb}, P_{eb}), where the discounted price P_{eb} is available to the first N_{eb} drivers that arrive prior to a_{eb}. This definition does not require the constraints given by N_{eb} and a_{eb} to be binding and so it incorporates the cases where either N_{eb} and a_{eb} is large, such that it is only the number of early birds or the latest arrival time of early birds that is constrained. Denote by $[e_0, e_1]$ the interval during which the early birds travel.

Drivers who do not receive the early bird special, we label regular drivers and we carry forward all previous notation to them: regular drivers pay the regular parking fee π, they travel during $[b_0, b_1]$ and b_* is the time after t_* where $v(b_0) = v(b_*)$. The following theorem characterizes welfare optimum under a parking fee combined with an early bird special. The welfare measure is again the sum of driver utility and parking fee revenues, which is simply the scheduling utility achieved.

Theorem 4 *Under the socially optimal combination of a regular parking fee π with an early bird special (N_{eb}, a_{eb}, P_{eb}), capacity is fully utilized throughout a period of length N/ψ, where $b_0 = e_0 + N_{eb}/\psi$ and $b_1 = e_0 + N/\psi$. The time-varying parking fee is*

$$\pi(t) = \begin{cases} 0, & t \in [b_0, b_*] \\ -v'(t) & t \in]b_*, b_1] \end{cases}.$$

There is queueing during $[b_0, b_*]$ and no queue during $[b_*, b_1]$. Departures begin later than in unregulated equilibrium such that $v(e_0) > v(b_1)$. The early bird charge lies between the total parking fees paid by the first and last regular drivers $P(b_1) < P_{eb} < P(b_0)$.

Figure 5 illustrates the social optimum for the general case. Evaluating the first order conditions for social optimum for the combination of a time-varying parking fee with an early bird special in the case of linear scheduling preferences as discussed in section 7 leads to $e_0 = -\frac{\gamma}{2}\frac{\beta+2\gamma}{(\beta+\gamma)^2}\frac{N}{\psi}$ and $b_0 = \frac{\gamma}{\beta+2\gamma}e_0$, such that $b_0 - e_0 = \frac{1}{2}\frac{\gamma}{\beta+\gamma}\frac{N}{\psi}$ and the optimal share of early birds out of all drivers is $\frac{1}{2}\frac{\gamma}{\beta+\gamma}$. With γ/β in the range $[2, 4]$, this share lies in the range $[0.33, 0.40]$ and it is always

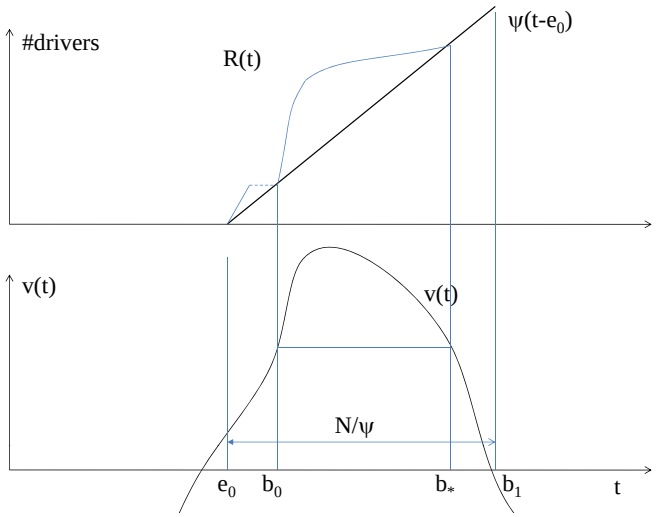

Figure 5: Social optimum with early bird special

smaller than $1/2$.

10 The evening commute

The analysis so far has concerned the morning commute, but with minor modifications it applies to the evening commute as well. This section will show that most conclusions carry more or less directly over from the morning to the evening commute.

Recall first that the analysis of the morning commute ignored any interaction with the evening commute, which could occur, e.g., through the duration of the period at work. This simplification greatly facilitates analysis and will be retained in the analysis of the evening commute.

Our general specification of scheduling preferences treats the departure time and the arrival time symmetrically, so it is not specific to the morning commute,

and applies equally well to the evening. We may consider scheduling preferences that are specific to the evening commute with t_* now being the preferred time of instantaneous transfer from work to home.

The treatment of congestion can also be exactly the same in the two commutes. Hence the evening no-policy equilibrium and the optimal time-varying toll exactly parallel those of the morning.

The difference is in the effect of a parking fee paid at the origin of the trip rather than at the destination. The parking fee is charged at the work place. Hence it creates an incentive to reduce the time spent at the workplace. This is equally true in both commutes. In the morning, the parking fee decreases with later departure (from home), while in the evening the parking fee increases with later departure (from work). This reversal has the effect of reversing the order of the two distinct intervals under the socially optimal parking fee. Recall that in the morning social optimum, there is first an interval of queueing, where the parking fee rate is zero, this is followed by an interval where the parking fee rate is $-v'$ and where there is no queue. In the evening social optimum, the evening parking fee rate is first equal to v' during an interval and this maintains the departure rate from work at the bottleneck capacity such that a queue does not arise. Later, in the evening, the parking fee rate is zero and a queueing interval occurs.

Early birds or drivers with private parking are not affected, they have no incentive to depart early and will depart during the period when the parking fee rate is zero. Thus the conclusions for the morning commute regarding drivers with private parking carry over to these cases.

11 Morning and evening commutes integrated

This section considers the morning and the evening commutes simultaneously and shows that interaction between commutes can imply that a parking fee can be designed to remove queueing completely. The parking fee is still restricted to be positive at any time so a parking fee in the morning can only reduce queueing in the morning but not remove it; similarly a parking fee in the evening can only reduce queueing in the evening. But it is possible to exploit interaction between the two commutes that occurs through the length of the time spent at work. Then the morning parking fee affects not only the morning commute but also the evening commute through the length of the working day; similarly a parking fee during the evening commute will affect the morning commute. Somewhat surprisingly, queueing can then be removed in both commutes simultaneously.

Consider drivers who commute to and from work. In the morning they pass through a bottleneck with capacity ψ_m, in the evening they pass through a bottleneck with capacity ψ_e and the two capacities may be different. The departure time from home in the morning is denoted t_m, departures begin at time c_m and cumulative departures in the morning are denoted R_m. Capacity will always be fully utilized during the commute such that $c_m + \frac{R_m(t_m)}{\psi_m}$ is the arrival time at work. The evening commute from home to work is denoted similarly with subscripts e.

We impose more structure on utility than we have before in this paper. In particular we assume that utility is separable in utility achieved at home at rate h_m prior to departure, utility achieved at home at rate h_e after returning home in the evening and utility achieved associated with the duration at work Γ.[10] Define then

[10]This assumes that workers can decide how much time to spend at work on any given day. An alternative would be to assume a fixed duration at work. This would however have the implication

the money-metric utility function

$$\begin{aligned} u\left(t_m, t_e\right) &= \int_0^{t_m} h_m\left(s\right) ds + \int_{c_e + \frac{R_e(t_e)}{\psi_e}}^{0} h_e\left(s\right) ds \\ &\quad + \Gamma\left(t_e - c_m - \frac{R_m\left(t_m\right)}{\psi_m}\right) - \int_{c_m + \frac{R_m(t_m)}{\psi_m}}^{t_e} \pi\left(s\right) ds. \end{aligned}$$

If it were the case that $\Gamma'' = 0$, then the utility function would be additively separable into a part depending only on t_m and another part depending only on t_e. In this case the morning and evening commutes could be analysed separately and we would be back in the situation from previous sections. So we require that $\Gamma > 0, \Gamma' > 0, \Gamma'' < 0$. Moreover, utility rates h_m, h_e satisfy $h_m, h_e > 0, h_m' < 0 < h_e'$. In order to guarantee existence of equilbrium it is sufficient (but not necessary) to assume that there is a point in time where $h_m\left(t\right) = h_e\left(t\right) < \Gamma'\left(0\right)$ and that h_m, h_e, Γ' all range from 0 to ∞. Parking is charged at the positive time-varying rate $\pi\left(\cdot\right)$ during the time spent at work.

It is clear from the previous analysis and for the same reasons as before that there are two commuting intervals in equilibrium, that capacity is fully utilized during these intervals if the parking fee is not too high, and that in each commute the queue is exactly gone at the time of the last departure. We assume that utility is such that the commuting intervals do not overlap. The equilibrium departure rates can be found from the first order conditions for utility maximization. The next lemma establishes that drivers pass the bottleneck in the same sequence in the two commutes. The lemma also states some inequalities that hold in equilibrium since

that the departure rate from work would be the same as the arrival rate at work, and this is at most a constant ψ_m. Then if $\psi_m < \psi_e$ there would never be queue in the evening or if $\psi_m > \psi_e$ there would be an increasing queue at all departure times from work where capacity ψ_m is utilized. Both implications seem strange.

the first driver in either commute will not prefer to depart earlier and that the last driver in either commute will not prefer to depart later, these are clearly necessary conditions for equilibrium to occur.

Lemma 1 *Drivers depart in the same sequence in the two commutes. There is a range of equilibria, determined by initial departure times c_m and c_e. The following inequalities hold in equilibrium:*

$$\begin{aligned} h_m(c_m) &\geq \Gamma'(c_e - c_m) - \pi(c_m) && (8)\\ h_m\left(c_m + \frac{N}{\psi_m}\right) &\leq \Gamma'\left(c_e + \frac{N}{\psi_e} - c_m - \frac{N}{\psi_m}\right) - \pi\left(c_m + \frac{N}{\psi_m}\right) && (9)\\ h_e(c_e) &\leq \Gamma'(c_e - c_m) - \pi(c_e) && (10)\\ h_e\left(c_e + \frac{N}{\psi_e}\right) &\geq \Gamma'\left(c_e + \frac{N}{\psi_e} - c_m - \frac{N}{\psi_m}\right) - \pi\left(c_e + \frac{N}{\psi_e}\right). && (11) \end{aligned}$$

The equilibrium with equality in (8) and (11) is Pareto dominant.

The lemma shows that a range of equilibria are possible. In the absence of queueing, the first driver would prefer to depart later from home. Likewise the last traveler would like to leave earlier from work if there were no queue. All drivers achieve the same utility in equilibrium. Therefore welfare is maximal if the equilibrium is the one with equality in (8) and (11). The next theorem establishes that it is possible to construct a parking fee that implements the Pareto dominant equilibrium such that there is no queueing in either commute.

Theorem 5 *Let times c_m and $c_e > c_m + N/\psi_m$ be given and define the function*

$$f(t_m) = \frac{\psi_m}{\psi_e}(t_m - c_m) + c_e, t_m \in [c_m, c_m + N/\psi_m]. \qquad (12)$$

Assume that c_m, c_e satisfy the following conditions:

$$\begin{aligned} h_m(c_m) &= \Gamma'(c_e - c_m) && (13)\\ h_e\left(c_e + \frac{N}{\psi_e}\right) &= \Gamma'\left(c_e + \frac{N}{\psi_e} - c_m - \frac{N}{\psi_m}\right) && (14)\\ \max\{h_m(t), h_e(f(t))\} &\leq \Gamma'(f(t) - t), t \in [c_m, c_m + N/\psi_m]. && (15) \end{aligned}$$

The following parking fee removes queueing completely and implements the unique Pareto optimal equilibrium:

$$\pi(t) = \begin{cases} \Gamma'(f(t) - t) - h_m(t) & c_m \leq t < c_m + N/\psi_m \\ \Gamma'(t - f^{-1}(t)) - h_e(t) & c_e \leq t < c_e + N/\psi_e \\ 0 & otherwise. \end{cases}$$

The parking fee of the theorem implements a situation where the first commute takes place during $[c_m, c_m + N/\psi_m]$ with departures at the capacity rate ψ_m. The definition (12) ensures that if drivers depart at the capacity rate ψ_m during the first commute, then they depart at the capacity rate ψ_e during $[c_e, c_e + N/\psi_e]$. Conditions (13-15) ensure that the parking fee rate is always positive and that $\pi(c_m) = \pi(c_e + N/\psi_e) = 0$. The equilibrium conditions in Lemma 1 are all satisfied with equality.

Compared to a situation with no parking fee and first departures still at c_m and c_e, the welfare gain from the parking fee of the theorem is total parking fee payment during the two commutes. The parking fee during $[c_m, c_e + N/\psi_e]$ when all are at work is set to zero in the theorem but can be larger provided the equilibrium conditions are not affected. The parking fee revenue during this period does then not affect behavior (as we assume fixed demand) and does hence not contribute to

any change in welfare.

12 Conclusion

This paper has analyzed the potential efficiency gains that may be realised through retiming of commuting trips due to a time-varying parking fee charged at a positive rate at the workplace. At the social optimum, the commute to work is divided into two distinct intervals by the optimal parking fee. During the first interval, parking is free and there is queueing. During the second interval, parking is charged at a time-varying rate such that there is no queue while capacity remains fully utilized. The sequence of these two periods is reversed from the morning to the evening commute. Parking fees create an incentive to reduce the length of time spent at work.

With private parking, a group of drivers cannot be charged for parking. It turns out not to matter for equilibrium departure time outcomes for the optimal charge, provided the drivers who cannot be charged are few enough to fit within the congested part of the commute. It is thus possible to exempt a group of drivers from paying the parking fee without sacrificing the welfare gains that can be achieved. Early bird specials may be designed to increase efficiency even further.

The analysis up to this point has treated the morning and evening commutes separately. During either commute, a parking fee can reduce congestion but not remove it. When there is interaction between the commutes through the duration of time spent at work then it is possible to affect the evening commute through a parking fee during the morning and vice versa. The paper has exhibited a case where it is then possible to utilize the interaction to remove congestion completely

during both commutes through a parking fee.

It is an essential characteristic of parking fees considered in the paper that the total parking fee payment is decreasing as a function of the arrival time at work in the morning and increasing as a function of the departure time from work in the evening. This restriction leads to results that differ from the case of a time-varying toll. If it were possible to charge for parking at a negative rate, then any time-varying toll could be replicated and the well-known analysis of such a toll could be applied.

It is straightforward to extend the results of this paper to the case of elastic demand. A way to proceed is to let aggregate demand depend on the average utility obtained in equilibrium. The optimal toll can then be obtained by fixing $P(b_1)$, which amounts to adding a fixed component to the parking fee. If $P(b_1) = -Nv'(b_1)\frac{\partial b_1}{\partial N}$ then the marginal benefit of adding a car equals the marginal cost. In this way the model can be extended to deal with externalities including e.g. congestion cruising for a limited number of parking spaces and other congestion externalities.

The current analysis has focused on the interaction of a time-varying parking fee rate with congestion dynamics. We focus on the timing of parking and thus complement the earlier contributions discussed in the introduction that, simply put, consider where and for how long to park. Future research could seek to integrate these perspectives in a unified analysis. It would also be natural to seek to allow for heterogenous drivers, as has been done for the bottleneck model by Lindsey (2004) and recently van den Berg and Verhoef (2011).[11]

[11] With the dynamic bottleneck model, METROPOLIS, implemented for large networks, such complications could be envisaged. This will allow to test the robustness of our predictions for large scale networks (see de Palma et al., 1997).

References

Aliprantis, C. D. and Border, K. C. (2006) *Infinite Dimensional Analysis: A Hitchhiker's Guide* Vol. 3rd. Springer-Verlag Berlin Heidelberg.

Anderson, S. and de Palma, A. (2004) The economics of pricing parking *Journal of Urban Economics* **55**(1), 1–20.

Anderson, S. and de Palma, A. (2007) Parking in the city *Papers in Regional Science* **86**(4), 621–632.

Arnott, R. A., de Palma, A. and Lindsey, R. (1993) A structural model of peak-period congestion: A traffic bottleneck with elastic demand *American Economic Review* **83**(1), 161–179.

Arnott, R., de Palma, A. and Lindsey, R. (1991) A temporal and spatial equilibrium analysis of commuter parking *Journal of Public Economics* **45**(3), 301–335.

Arnott, R. and Inci, E. (2006) An integrated model of downtown parking and traffic congestion *Journal of Urban Economics* **60**(3), 418–442.

Arnott, R. and Rowse, J. (1999) Modeling Parking *Journal of Urban Economics* **45**(1), 97–124.

Arnott, R. and Rowse, J. (2009) Downtown parking in auto city *Regional Science and Urban Economics* **39**(1), 1–14.

Calthrop, E., Proost, S. and Van Dender, K. (2000) Parking policies and road pricing *Urban Studies* **37**(1), 63–76.

Daganzo, C. F. (2007) Urban gridlock: Macroscopic modeling and mitigation approaches *Transportation Research Part B: Methodological* **41**(1), 49–62.

De Borger, B. and Proost, S. (2010) A political economy model of road pricing.

De Borger, B. and Wuyts, B. (2009) Commuting, Transport Tax Reform and the Labour Market: Employer-paid Parking and the Relative Efficiency of Revenue Recycling Instruments *Urban Studies* **46**(1), 213–233.

de Palma, A. and Fosgerau, M. (2011a) Dynamic Traffic Modeling *in* A. de Palma, R. Lindsey, E. Quinet and R. Vickerman (eds), *A Handbook of Transport Economics* Edward Elgar.

de Palma, A. and Fosgerau, M. (2011b) Random queues and risk averse users. DTU Transport.

de Palma, A. and Lindsey, R. (2002) Comparison of Morning and Evening Commutes in the Vickrey Bottleneck Model *Transportation Research Record: Journal of the Transportation Research Board* **1807**(1), 26–33.

de Palma, A., Marchal, F. and Nesterov, Y. (1997) METROPOLIS - Modular System for Dynamic Traffic Simulation *Transportation Research Record* **1607**, 178–184.

Douglas, R. W. (1975) A parking model - the effect of supply on demand *American Economist* **19**(1), 85–86.

Fosgerau, M. (2011) How a fast lane may replace a congestion toll *Transportation Research Part B* **45**(6), 845–851.

Geroliminis, N. and Daganzo, C. F. (2008) Existence of urban-scale macroscopic fundamental diagrams: Some experimental findings *Transportation Research Part B: Methodological* **42**(9), 759–770.

Glazer, A. and Niskanen, E. (1992) Parking fees and congestion *Regional Science and Urban Economics* **22**(1), 123–132.

Lindsey, R. (2004) Existence, Uniqueness, and Trip Cost Function Properties of User Equilibrium in the Bottleneck Model with Multiple User Classes *Transportation Science* **38**(3), 293–314.

Proost, S. and Van Dender, K. (2008) Optimal urban transport pricing in the presence of congestion, economies of density and costly public funds *Transportation Research Part A: Policy and Practice* **42**(9), 1220–1230.

Qian, Z. S., Xiao, F. E. and Zhang, H. (2012) Managing morning commute traffic with parking *Transportation Research Part B: Methodological* **46**(7), 894–916.

Schrank, D., Lomax, T. and Eisele, B. (2011) TTI's 2011 Urban Mobility Report *Technical report* Texas Transportation Institute, Texas A&M University College Station.

Shoup, D. (2005) *The High Cost of Free Parking* Planners Press, American Planning Association Chicago, Illinois, Washington D.C.

Small, K. (1982) The scheduling of Consumer Activities: Work Trips *American Economic Review* **72**(3), 467–479.

van den Berg, V. and Verhoef, E. T. (2011) Winning or losing from dynamic bottleneck congestion pricing?: The distributional effects of road pricing with het-

erogeneity in values of time and schedule delay *Journal of Public Economics* **95**(7-8), 983–992.

Van Ommeren, J., Wentink, D. and Dekkers, J. (2011) The real price of parking policy *Journal of Urban Economics* **70**(1), 25–31.

Verhoef, E., Nijkamp, P. and Rietveld, P. (1995) The economics of regulatory parking policies: The (IM)possibilities of parking policies in traffic regulation *Transportation Research Part A: Policy and Practice* **29**(2), 141–156.

Vickrey, W. (1954) The economizing of curb parking space. (November), 62–67. Reprinted in Journal of Urban Economics 36, 1994, 4265.

Vickrey, W. (1969) Congestion theory and transport investment *American Economic Review* **59**(2), 251–261.

Vickrey, W. S. (1973) Pricing, metering, and efficiently using urban transportation facilities *Highway Research Record* **476**, 36–48.

Victoria Transport Policy Institute (2012) TDM Encyclopedia, Parking Pricing.

Zhang, X., Yang, H. and Huang, H.-J. (2005) Integrated scheduling of daily work activities and morning-evening commutes with bottleneck congestion *Transportation Research Part A* **39**(1), 41–60.

A Proofs

Proof of Theorem 2. The first part of the theorem has already been established. It remains to determine the welfare maximizing value of Δ. Compute the derivative of W as

$$W'(\Delta) = \psi(b'_* - b'_0) v(b_0) + \psi(b_* - b_0) v'(b_0) b'_0 + \psi(v(b_1) b'_1 - v(b_*) b'_*).$$

Use that $v(b_0) = v(b_*)$, $b'_1 = b'_0$, and $\Delta = v(b_0) - v(b_1)$ to reduce this expression to

$$W'(\Delta) = [(b_* - b_0) v'(b_0) - \Delta] \psi b'_0,$$

and note that this is zero if and only if $\Delta = (b_* - b_0) v'(b_0)$. Next use that $1 = v'(b_0) b'_0 - v'(b_1) b'_1$ and $b'_1 = b'_0$ to find that $b'_0 = (v'(b_0) - v'(b_1))^{-1} > 0$. Note that

$$W'(0) = [(b_* - b_0) v'(b_0)] \psi b'_0 > 0$$

and that

$$W'(\Delta^*) = -\Delta^* \psi b'_0 < 0,$$

since $b_0 = b_*$ at $\Delta = \Delta^*$. Then there is at least one value of Δ between 0 and Δ^* with $W'(\Delta) = 0$. Evaluate next the second derivative of W at a point with $W'(\Delta) = 0$:

$$\begin{aligned} W''(\Delta) &= [(b'_* - b'_0) v'(b_0) + (b_* - b_0) v''(b_0) b'_0 - 1] \psi b'_0 \\ &+ [(b_* - b_0) v'(b_0) - \Delta] \psi b''_0 \\ &= [(b'_* - b'_0) v'(b_0) + (b_* - b_0) v''(b_0) b'_0 - 1] \psi b'_0 \end{aligned}$$

$$= \left[\frac{v'(b_0) - v'(b_*)}{v'(b_*)} v'(b_0) b_0' + (b_* - b_0) v''(b_0) b_0' - 1\right] \psi b_0',$$

where the last equality follows upon noting that $v'(b_0)b_0' = v'(b_*)b_*'$. This is negative if and only if

$$\frac{v'(b_0) - v'(b_*)}{v'(b_*)} v'(b_0) b_0' + (b_* - b_0) v''(b_0) b_0' < 1.$$

But this inequality holds since $v'(b_*) < 0$ and v is concave. Thus $W'(\Delta) = 0$ implies that $W''(\Delta) < 0$ and hence that W is quasiconcave on the interval $[0, \Delta^*]$ such that W has a unique maximum there. It is straightforward to verify that this maximum is global. ■

Proof of Theorem 3. Given the assumptions of Theorem 1, all departures will take place within the interval $[b_0, b_1]$ in Nash equilibrium. Now, $v(b_0) = u\left(b_0, \frac{R(b_0)}{\psi} + b_0\right) > u\left(b_1, \frac{R(b_1)}{\psi} + b_0\right) = v(b_1)$, where the inequality follows since the last driver pays a strictly smaller parking fee than the first but achieves the same utility. Moreover, $U^u \equiv u\left(t, \frac{R(t)}{\psi} + b_0\right), t \in S^u$ is constant, which requires that there is queue almost always during S^u. Equilibrium similarly requires that $U^c \equiv u\left(t, \frac{R(t)}{\psi} + b_0\right) - P(t), t \in S^c$ is constant. Thus $u\left(t, \frac{R(t)}{\psi} + b_0\right)$ is strictly decreasing on points of S^c where $\pi(t) > 0$. These conditions imply that all uncharged drivers obtain utility $v(b_0)$. Therefore they must all depart in the interval $[b_0, b_*]$, where b_* is defined by the equation $v(b_0) = v(b_*)$, which implies that $b_* < b_1$ by quasiconcavity of v. ■

Proof of Proposition 1. Given $\Delta = P(b_0) - P(b_1)$, with $0 < \Delta < v(t_*) -$

$v(a_0) = \frac{N}{\psi}\frac{\beta\gamma}{\beta+\gamma}$ and $P(b_1) = 0$, it is straightforward to find that

$$b_0 = \frac{\Delta - \gamma\frac{N}{\psi}}{\beta+\gamma}, b_* = -\frac{\beta}{\gamma}\frac{\Delta - \gamma\frac{N}{\psi}}{\beta+\gamma}, b_1 = \frac{\Delta + \beta\frac{N}{\psi}}{\beta+\gamma}.$$

Then the welfare given Δ is

$$\begin{aligned} W(\Delta) &= Nv(b_0) - \psi(b_1 - b_*)\Delta/2 \\ &= N\beta\frac{\Delta - \gamma\frac{N}{\psi}}{\beta+\gamma} - \psi\left(\frac{\Delta + \beta\frac{N}{\psi}}{\beta+\gamma} + \frac{\beta}{\gamma}\frac{\Delta - \gamma\frac{N}{\psi}}{\beta+\gamma}\right)\frac{\Delta}{2} \\ &= \frac{1}{\beta+\gamma}\left(-\beta\gamma\frac{N^2}{\psi} + \beta\Delta N - \psi\frac{\beta+\gamma}{\gamma}\frac{\Delta^2}{2}\right). \end{aligned}$$

This is maximal when

$$\Delta = \frac{\beta\gamma}{\beta+\gamma}\frac{N}{\psi}.$$

In this case

$$b_0 = \frac{-\gamma^2}{(\beta+\gamma)^2}\frac{N}{\psi}, b_* = \frac{\beta\gamma}{(\beta+\gamma)^2}\frac{N}{\psi}, b_1 = \frac{\beta^2+2\beta\gamma}{(\beta+\gamma)^2}\frac{N}{\psi}.$$

The optimal time-varying toll leads to a welfare gain of $\frac{N^2}{\psi}\frac{\beta\gamma}{2(\beta+\gamma)}$. ■

Proof of Theorem 4. Clearly, early birds depart before other drivers during $[e_0, e_1]$ where $e_1 < t_*$. They pay the same price for parking and will therefore queue, departing at the rate $\rho_{eb}(t) > \psi$,with $R_{eb}(e_1) = N_{eb}$ and the last arrival time being $e_0 + \frac{N_{eb}}{\psi}$. For other drivers, it is optimal that they are charged according to a fee as in section 6 where there is first an interval $[b_0, b_*]$ of arrival times where the parking fee rate is zero, there is queueing and $v(b_0) = v(b_*)$, next there is an interval $[b_*, b_1]$ of arrival times with no queueing and a parking fee rate that is

$\pi(t) = -v'(t)$. We recall that $b_0 \leq t_* \leq b_* < b_1$. It is also clear that capacity should be fully utilized during the commute. This requires that the last arrival time of the early birds is the same as the first arrival time of the ordinary drivers $e_0 + \frac{N_{eb}}{\psi} = b_0$. All drivers pass the bottleneck during $[e_0, b_1]$, so $b_1 = e_0 + N/\psi$. Thus the timing of departures is determined by e_0 and b_0. The difference between the parking fees P_{eb} and P is then also determined since all drivers achieve the same utility in equilibrium.

Welfare is

$$W = \psi \cdot (b_0 - e_0)\, v(e_0) + \psi \cdot (b_* - b_0)\, v(b_0) + \psi \int_{b_*}^{b_1} v(t)\, dt,$$

which is composed of $\psi \cdot (b_0 - e_0)$ early birds achieving scheduling utility $v(e_0)$, $\psi \cdot (b_* - b_0)$ ordinary drivers achieving scheduling utility $v(b_0)$ and the remaining $\psi \cdot (b_1 - b_0)$ achieving scheduling utility $v(t)$. The timing of departures is chosen through e_0 and b_0 to optimize welfare with first order conditions (when $b_0 < t_* < b_*$)

$$\begin{aligned} v(e_0) &= (b_0 - e_0)\, v'(e_0) + v(b_1), \\ v(e_0) &= v(b_0) - (b_* - b_0)\, v'(b_0). \end{aligned}$$

Now $v'(e_0), v'(b_0) > 0$ such that $v(b_1) < v(e_0) < v(b_0)$. Utilities are equal in equilibrium so $P(b_1) < P_{eb} < P(b_0)$.

A corner solution arises when $b_0 = t_* = b_*$. In that case only e_0 may vary and

has first order condition

$$v\left(e_0\right) = \left(t_* - e_0\right) v'\left(e_0\right) + v\left(b_1\right),$$

implying that again $v\left(b_1\right) < v\left(e_0\right) < v\left(b_0\right)$. ■

Proof of Lemma 1. The first order condition for the choice of departure time in the morning, given the departure time in the evening, is

$$0 = \frac{\partial u\left(t_m, t_e\right)}{\partial t_m} = h_m\left(t_m\right) - \left(\Gamma'\left(t_e - c_m - \frac{R_m\left(t_m\right)}{\psi_m}\right) - \pi\left(c_m + \frac{R_m\left(t_m\right)}{\psi_m}\right)\right)\frac{\rho_m\left(t_m\right)}{\psi_m}.$$

Observe that any t_m can only solve the first order condition for one value of t_2. The function $t_e\left(t_m\right)$ thus defined then is single-valued. By the Berge maximum theorem (Aliprantis and Border, 2006), t_e has compact graph and hence t_e is continuous. We take for granted that it is continuously differentiable. The second order condition requires that

$$\frac{\partial^2 u\left(t_m, t_e\right)}{\partial t_m^2} \leq 0.$$

Differentiating the first order condition with respect to t_m leads to

$$0 = \frac{\partial^2 u\left(t_m, t_e\right)}{\partial t_m^2} - \Gamma''\left(t_e - c_m - \frac{R_m\left(t_m\right)}{\psi_m}\right)\frac{\partial t_e}{\partial t_m}$$

and hence $\frac{\partial t_e}{\partial t_m} \geq 0$. It is possible to have $\frac{\partial t_e}{\partial t_m} = 0$ at points, but $\frac{\partial t_e}{\partial t_m} = 0$ cannot hold on any interval. If it did, then there would be a mass departure in the evening, which is ruled out in equilibrium (if a mass departure should occur, then it is always strictly utility increasing to postpone departure until immediately after the

mass departure). This shows that $\frac{\partial t_e}{\partial t_m} > 0$ almost everywhere.

The inequalities characterize equilibrium since they imply that the first driver in either commute will not prefer to depart earlier and that the last driver in either commute will not prefer to depart later. Equality of utility holds due to queueing. With equality in (8) and (11), the first driver would not have incentive to postpone departure if there were no queue. ■

Proof of Theorem 5. Let $R_m(t_m) = \psi_m(t_m - c_m)$ in the first commute and $R_e(t_e) = \psi_e(t_e - c_e)$ in the second. Then there is no queueing while capacity is fully utilized. Utility for a driver with departure times t_m and t_e is then

$$u(t_m, t_e) = \int_0^{t_m} h_m(s)\, ds + \int_{t_e}^{0} h_e(s)\, ds + \Gamma(t_e - t_m) - \int_{t_m}^{t_e} \pi(s)\, ds.$$

Consider a driver departing at time $t_m \in [c_m, c_m + N/\psi_m]$. Then the first order condition for the choice of the second departure time has only one solution, namely at $t_e = f(t_m)$ by the definition of π. Moreover, the second order condition is satisfied,

$$\begin{aligned}
\left.\frac{\partial^2 u(t_m, t)}{\partial t^2}\right|_{t=t_e} &= -h'(t_e) + \Gamma''(t_e - t_m)\left(1 - \frac{1}{f'(t_m)}\right) - \pi'(t_e) \\
&= -h'(t_e) + \Gamma''(t_e - t_m) - \left(\Gamma''\left(t_e - f^{-1}(t_e)\right)\left(1 - \frac{1}{f'(t_m)}\right) - h'(t_e)\right) \\
&= \frac{\Gamma''(t_e - t_m)}{f'(t_m)} < 0.
\end{aligned}$$

With the optimal choice of departure time from work, $t_e = f(t_m)$, utility is

constant over the interval $t_m \in [c_m, c_m + N/\psi_m]$, since

$$\frac{\partial u\left(t_m, f\left(t_m\right)\right)}{\partial t_m} = h_m\left(t_m\right) - \Gamma'\left(t_e - t_m\right) + \pi\left(t_m\right) + \left(-h_e\left(t_e\right) + \Gamma'\left(t_e - t_m\right) - \pi\left(t_e\right)\right) f'\left(t_m\right) = 0,$$

by the definition of π. Then the departure rates R_m, R_e defined above do in fact lead to equilibrium.

The equilibrium conditions in Lemma 1 are satisfied by construction of π. Conditions (8) and (11) are satisfied with equality, indicating that the Pareto dominant equilibrium is implemented. ■

V. FILES D'ATTENTES ALEATOIRES ET USAGERS AVERSES AU RISQUE

Mogens Fosgerau, Technical University of Denmark et Center for Transport Studies, Suède

André de Palma, Ecole Normale Supérieure de Cachan et Centre d'économie de la Sorbonne. CES, France

RESUME

Nous étudions un service avec demande de pointe. On s'intéresse au cas où la demande est supérieure à la capacité, de sorte qu'il y a une file d'attente devant le service. Ce service sujet à congestion correspond à une route, à un parc d'attraction, ou une rame de métro surchargée. Le nombre d'usagers pouvant être servis par unité de temps est donné. Le temps de service dépend à la fois du temps d'arrivée dans la file d'attente, et aussi de la discipline de service. On considère deux cas limites.

Le premier est celui de la file d'attente déterministe. Dans ce cas, le temps d'attente est égale au nombre de personnes dans la file à l'arrivée divisé par la capacité de la file.

Dans l'autre cas limite, l'ordre dans la file est aléatoire, de sorte qu'à tout instant, chaque usager possède la même probabilité d'être servi. Le système doit satisfaire la contrainte globale selon laquelle le nombre d'usagers servis par unité de temps est égal à la capacité du service, donnée.

Nous analysons enfin les cas intermédiaires entre ces deux cas limites et en proposant une modélisation.

On suppose que les préférences des usagers sont décrits par une fonction d'utilité concave (ou linéaire par morceau) de sorte que les usagers sont averses au risque. Nous introduisons une condition dite d'absence de file d'attente résiduelle. Selon cette condition, il n'y a pas de file d'attente lorsque le dernier usager entré arrive au service. On montre que cette condition permet de garantir l'existence d'un équilibre, dans les deux cas limite et dans les cas intermédiaires. Cette condition permet aussi de calculer facilement les coûts d'équilibre et d'optimum sous les différents régimes, ainsi que la dynamique de la congestion.

Mots clé : file d'attente, aversion au risque, arrivée endogènes, service avec demande de pointe

Codes JEL: D00; D80

Random queues and risk averse users*

André de Palma† Mogens Fosgerau‡

October, 2012

Abstract

We analyse Nash equilibrium in time of use of a congested facility. Users are risk averse with general concave utility. Queues are subject to varying degrees of random sorting, ranging from strict queue priority to a completely random queue. We defin the key "no residual queue" property, which holds when there is no queue at the time the last user arrives at the queue, and prove that this property holds in equilibrium under all queueing regimes considered. The no residual queue property leads to simple results concerning the equilibrium utility of users and the timing of the queue.

Keywords: Congestion; Queuing; Risk aversion; Endogenous arrivals.
JEL codes: D00; D80

*This research is part of the SURPRICE project as well as of PREDIT-ADEME : TARIFICATION DES TRANSPORTS INDIVIDUELS ET COLLECTIFS A PARIS DYNAMIQUE DE L'ACCEPTABILITE and PREDIT: SCHEDULING, TRIP TIMING AND SCHEDULING PREFERENCES. We are grateful to Robin Lindsey, Katrine Hjorth, Hugo Harari-Kermadec, Søren Feodor Nielsen, Ken Small and seminar participants at the University of Copenhagen and at the Swedish Royal Institute of Technology for comments. Mogens Fosgerau is supported by the Danish Social Science Research Council. A special thanks is due to Richard Arnott, who gave as a number of very useful comments.

†École Normale Supérieure de Cachan, andre.depalma@ens-cachan.fr.

‡Corresponding author: Technical University of Denmark & Centre for Transport Studies, Sweden. mf@transport.dtu.dk

1 Introduction

We generalise the Vickrey (1969) analysis of bottleneck congestion to allow for random queue sorting as well as more general scheduling preferences. The paper shows that the fundamental insights of Vickrey remain valid in these circumstances. In spite of users being risk averse, random queue sorting turns out to play no role for the properties of equilibrium that are relevant for regulation of congestion.

Enormous amounts of time are lost queueing. Just for private transportation, the cost of congestion in Europe and the US is equivalent to more than 1 percent of GDP (International Transport Forum, 2007; Texas Transportation Institute, 2007) and unpriced congestion leads to excess urban sprawl (Arnott, 1979). Dynamic models of traffic congestion are reviewed in de Palma and Fosgerau (2011). Congestion arises not only on roads. Queues occur regularly also in supermarkets, banks, public offices, restaurants (Becker, 1991), movie theatres, concert ticket sales, at ski lifts (Barro and Romer, 1987) and toll road booths, in airports (Daniel, 1995), computer systems, communications systems, web services, call centers, and many other systems. Queueing is also relevant for understanding competitive markets, where queueing plays a role in allocating goods among consumers and trade from firms is congestible (Sattinger, 2002). So it is clearly important to understand queueing phenomena.

Economic analyses of congestion mostly assume strict first-in-first-out (FIFO) queue discipline, whereby the order of arrival at the queue is preserved. Many real queues, however, involve an element of random sorting. An extreme case is a pure random queue.[1] An example is a (virtual) queue to get through on a busy telephone line (de Palma and Arnott, 1989), where every person present in the queue at a given time has the same probability of being served as any other person in the queue, regardless of how long each has been in the queue. Other queues also involve random queue sorting. There are random opportunities for overtaking on roads; in a supermarket, FIFO applies to individual checkout lines, but not to the supermarket checkout system as a whole (Blanc, 2009); also queueing for public transport is often not strictly FIFO (Yoshida, 2008). In general, we may think that strict FIFO rarely occurs. It is thus of interest to determine the properties of queues that are not strictly FIFO.[2]

The economic literature has previously paid attention to the properties of user equilibrium in queues with strict queue priority using the seminal Vickrey (1969)

[1] It is also possible to conceive of queues with a queue manager. In this case, a last-in-first-out queue may be considered an opposite of a FIFO queue (Hassin, 1985).

[2] Arnott, de Palma and Lindsey (1996) and (Arnott, de Palma and Lindsey, 1999) analyze a situation in which capacity varies randomly from day to day, while the queue retains the FIFO property.

bottleneck model. This model offers many insights that are central to the understanding of congested demand peaks. Arnott, de Palma and Lindsey (1993) summarise a number of these. In the Vickrey model, users arrive at a bottleneck where they wait in a FIFO queue until they are served by the bottleneck. The bottleneck serves users at a fixed rate. A continuum of users choose their time of arrival at bottleneck into the queue to minimise a scheduling cost, which is linear in time spent in the queue, time early and time late at the destination. The time-varying arrival rate at the bottleneck is then determined endogenously in response to the evolution of the queue. The model is closed by assuming Nash equilibrium.[3]

We extend the Vickrey model in two ways: first by allowing for random queue sorting, and second by allowing users to have general concave utility depending on duration in the queue as well as on time of exit from the queue. Random queue sorting causes randomness in outcomes and the concavity of utility implies that users are risk averse.

We then introduce the no residual queue (NRQ) property for a queue with a general random sorting mechanism. A residual queue is a queue that remains at the time of arrival at the bottleneck of the last user. The NRQ property is said to hold when the queue has vanished at the time of the last arrival. By definition, the equilibrium utilities of the first and the last user are equal. The NRQ property is then sufficient to establish the equilibrium time interval of arrivals. A number of useful results follow. In particular, we determine the equilibrium utility and the marginal utility of adding users under Nash equilibrium. This is the information that is needed in order to determine the optimal capacity provision as well the optimal constant toll.

The basic insight is then that it is the NRQ property that underlies the elegance of the Vickrey analysis of congestion. When the NRQ property holds, it does not matter that the queue is subject to random sorting. Remarkably, the optimal capacity, the optimal constant toll as well as the optimal time varying toll are unaffected by random queue sorting.

So it is of interest to establish when the NRQ property holds. We identify a condition on scheduling preferences that is sufficient for the NRQ property under any degree of random queue sorting. It turns out to be sufficient that users must be always willing to arrive one minute later in exchange for spending one minute less in the queue. This condition cannot be relaxed in general.

We also show that the optimal time varying toll is also not affected by random queue sorting, since there is no queue under the optimal time varying toll. This result holds regardless of whether the NRQ property holds in no toll equilibrium.

The paper is organised as follows. Section 2 presents the general framework,

[3]The operations research literature generally considers the arrival rate as exogenous, perhaps allowing the user to balk when he meets a long queue (Naor, 1969; Knudsen, 1972).

introduces the NRQ property, and derives the results that follow from this property.

The remainder of the paper is devoted to establishing the NRQ property under various degrees of random queue sorting. First, Section 3 reviews and generalises the standard case of *strict queue priority* and establishes that the NRQ property holds here. Next, Section 4 considers the oppposite case of *no queue priority* where users to be served are chosen completely at random from the queue. We establish also the NRQ property for this case given the above condition on preferences.

Section 5 considers the intermediate case, which we refer to as *loose queue priority*. Under this regime, the probability of being served at time t, conditional on being in the queue at time t, increases with the time spent in the queue. We show that the above condition on marginal utilities is again sufficient to guarantee the NRQ property to hold in general when queue priority is loose. Some concluding remarks are provided in Section 6.

2 Model specification

Consider N users treated as a continuum. They must all pass through a bottleneck which has a capacity of ψ users per time unit. Users arrive at the bottleneck at the back of the queue at the locally bounded time dependent rate $\rho(a) \geq 0$ during the interval $[t_0, t_1]$, where t_0 and t_1 are the minimum and the maximum of the support of ρ. The cumulative arrival rate up to time a is denoted by $R(a) = \int_{t_0}^{a} \rho(s)\, ds$, and $R(\cdot)$ is continuous since $\rho(\cdot)$ is locally bounded. Furthermore, $R(\cdot)$ is differentiable at all points of continuity of $\rho(\cdot)$. Users enter a vertical queue of length $Q(a)$ at time a, which represents the number of users who have arrived at the entrance of the bottleneck but not yet exited. The queue length evolves according to[4]

$$Q(a) = R(a) - \int_{t_0}^{a} \left[\psi 1_{\{Q(s)>0\}} + \min(\psi, \rho(s))\, 1_{\{Q(s)=0\}}\right] ds, \tag{1}$$

so $Q(\cdot)$ is continuous and also differentiable at points of continuity of $\rho(\cdot)$. Denote the minimum and the maximum of the support of the queue length $Q(\cdot)$ as τ_0 and τ_1.

The last user exits the queue at time τ_1. This implies that $\tau_1 \geq t_1$. If $Q(t_1) = 0$, then $\tau_1 = t_1$. If $Q(t_1) > 0$, we say that there is a *residual queue* at time t_1. In this case, τ_1 is given by $Q(t_1) = \psi(\tau_1 - t_1)$, since the queue length at time $t \in [t_1, \tau_1[$ is strictly positive if $Q(t_1) > 0$.

[4] $1_{\{\cdot\}}$ is the indicator function for the event in curly brackets.

We shall consider various queueing regimes. At one extreme we have the *strict queue priority* case, considered by Vickrey (1969), where the queue obeys the first-in-first-out principle (FIFO). At the other extreme we have the *no queue priority* case, where the user to exit at each instant is chosen completely at random from the queue. Therefore the probability of exit from the queue at some instant is the same for all users present in the queue and does not depend on how much time each has spent in the queue. In between these two cases, we have the *loose queue priority* case. In this case, users who are in the queue in a given instant have a higher probability of exit if they have spent more time in the queue.

We formalise these cases below through the conditional density of exit times $f(t|a)$, which describes the probability of exit at time t conditional on arrival at the bottleneck at time $a \leq t$. This conditional density depends on the arrival rate $\rho(\cdot)$, but it is exogenous from the perspective of a single atomistic user. In all cases, except the strict queue priority case that is treated separately, we assume that $f(t|a)$ is differentiable as a function of a.

A user arrives at the bottleneck at time a and exits at time t with $a \leq t$, such that his duration in the queue is $d = t - a$. The arrival time is chosen by the user while the exit time is determined by the queue. He has a preferred exit time t^*. Utility is associated with the duration in the queue and the deviation $t - t^*$ of the exit time from the preferred exit time. Assume homogenous users and write utility as $u(d, t - t^*)$. Utility is concave, has a unique maximum at $d = 0$ for any $t - t^*$ and a unique maximum at $t = t^*$ for any duration in the queue. Given any exit time, users strictly prefer zero duration in the queue to anything else, and given any duration in the queue, users strictly prefer exiting at the preferred time to anything else. With these assumptions, utility is strictly decreasing in d, strictly increasing in t for $t < t^*$ and strictly decreasing in t for $t > t^*$. We normalise $t^* = 0$ at no loss of generality.

Users choose their arrival time a to maximise their expected utility given by

$$E(u|a) = \int_a^\infty u(t - a, t) f(t|a) \, dt. \tag{2}$$

We specify the following assumptions concerning the utility function. Denote the partial derivatives of u with respect to duration and exit time as u_1 and u_2, respectively. We require first and second derivatives to exist, except $u_2(d, 0)$ which is not required to exist. Clearly, users who exit late are always willing to exit one minute earlier in exchange for spending one minute less in the queue. We require that also users who exit early are always willing willing to exit one minute earlier in exchange for spending one minute less in the queue. This first condition is assumed throughout the paper.

Condition 1 $u_1(d,t) + u_2(d,t) < 0$ *for all* $t < 0$.

We shall also have use for a second condition stating that users who exit late are always willing to exit one minute later in exchange for spending one minute less in the queue. For easy reference we shall call this the *acceptable lateness* condition. Clearly, users who exit early always satisfy the acceptable lateness condition. It is assumed where indicated.

Condition 2 *(Acceptable lateness)* $u_1(d,t) < u_2(d,t)$ *for all* $t > 0$.

We shall refer to the special case of linear utility, which is the case investigated by Vickrey (1969) and Arnott et al. (1993). This will be important for results and also helps in facilitating interpretation of results. The linear utility formulation is[5]

$$u(d,t) = -\alpha d - \beta t^- - \gamma t^+,$$

where the parameters α, β and γ are strictly positive. For the linear case, condition 1 states that $\beta < \alpha$, while the acceptable lateness condition 2 states that $\gamma < \alpha$. Yoshida (2008) summarises empirical evidence and concludes that both cases $\gamma < \alpha$ and $\gamma > \alpha$ are empirically relevant.

We consider Nash equilibrium in pure strategies as the benchmark for rational behavior.[6] The Nash equilibrium is defined by the requirement that, conditional on the actions of other users, no user has incentive to change his own action. With identical users, this requirement turns into the condition that the expected utility is constant and maximal over the times at which users arrive, i.e. over the support of ρ.

Below we shall briefly touch the issue of optimal tolling. For this we need to specify how a toll payment enters utility and a social welfare function with respect to which optimality is defined. We take any toll payment to be simply subtracted from utility, which then must be in monetary units. When expected utility is constant over users, we define a social welfare function as N times the equilibrium expected utility plus aggregate toll revenues.

In the strict queue priority case, the exit time is given deterministically as a function of the arrival time. We then require that utility is constant over all arrival times a with $\rho(a) > 0$.

In all other cases considered, exit time is random. The Nash condition implies that the expected utility is constant, i.e. $\frac{\partial E(u|a)}{\partial a} = 0$, for all a such that $\rho(a) > 0$.

[5] $x^+ = \max(x, 0)$, and $x = x^+ - x^-$.

[6] The equilibrium concept is discussed by Arnott et al. (1993).

This leads to the equation

$$-u\left(0,a\right)f\left(a|a\right)+\int_{a}^{\infty}\left[u\left(t-a,t\right)\frac{\partial f\left(t|a\right)}{\partial a}-u_{1}\left(t-a,t\right)f\left(t|a\right)\right]dt=0.$$

Recall that t_0 and t_1 are the times of the first and the last arrival. The following Lemma shows that in equilibrium the queue begins when the first user arrives at the bottleneck and that the queue ends at the earliest when the last user arrives.

Lemma 1 *The support of Q is a finite interval in Nash equilibrium, with $-\infty < t_0 = \tau_0 < 0$ and $0 < t_1 \leq \tau_1 < \infty$.*

All proofs are given in the appendix. We now introduce the no residual queue property.

Definition 1 *The no residual queue (NRQ) property holds if $\tau_1 \leq t_1$.*

The NRQ property ensures that $[t_0, t_1] = [\tau_0, \tau_1]$ in Nash equilibrium by Lemma 1. This means that the first and last users experience no queue, and hence that $u\left(0,t_0\right) = u\left(0,t_1\right)$. Moreover, all users are able to pass the bottleneck during $[t_0, t_1]$, which implies that $t_1 = t_0 + N/\psi$. These two observations pin down the equilibrium utility as shown in the following Proposition.

Proposition 1 *Consider Nash equilibrium where the NRQ property holds. Then the interval of arrival, $[t_0, t_1]$ with $t_0 < 0 < t_1$, is uniquely determined by $t_1 = t_0 + \frac{N}{\psi}$ and $u\left(0,t_0\right) = u\left(0, t_0 + \frac{N}{\psi}\right)$. The expected utility of any user is $u\left(0,t_0\right)$. The marginal change in expected utility from additional users is*

$$\frac{\partial E\left(u|a\right)}{\partial N} = \frac{1}{\psi}\frac{u_2\left(0,t_0\right)u_2\left(0,t_1\right)}{u_2\left(0,t_1\right)-u_2\left(0,t_0\right)} < 0, \tag{3}$$

which decreases in the number of users.

The preceding Proposition exhibits the central properties of the bottleneck model. In particular, the expected utility of any user is known as a function of the number of users, which makes it easy to derive the optimal capacity. If the number of users is allowed to be elastic, then Proposition 1 can be used to determine the optimal constant toll. Below we establish that the NRQ property holds under strict, loose and no queue priority and hence that Proposition 1 applies in all these regimes.

The optimal time varying toll eliminates queueing. Hence it is not affected by random queue sorting. This is formalised in the following Proposition, which is stated without proof.

Proposition 2 *The optimal time varying toll is*

$$[u(0,a) - u(0,t_0)]^+,$$

where t_0 is the first arrival time in Nash equilibrium under strict queue priority.

3 Strict queue priority

This is the case considered by Vickrey (1969) and Arnott et al. (1993) in the context of transportation and telecommunication, except for our more general formulation of user preferences. Users exit strictly in the order in which they arrive, hence exit time is a deterministic function of arrival time. A user arriving at time a is served at time $a + q(a)$, where $q(a) = Q(a)/\psi$. We have $q(a) = \frac{R(a)}{\psi} - (a - t_0)$, since there is always queue during $[t_0, t_1]$. Therefore

$$q'(a) = \frac{\rho(a)}{\psi} - 1. \tag{4}$$

The queue satisfies the NRQ property, since if the last user arrives at time t_1 when $Q(t_1) > 0$, then his exit time will be $\tau_1 > t_1$. This implies that he could postpone arrival until τ_1 to obtain zero duration in the queue while leaving the exit time unchanged, in contradiction of Nash equilibrium. We highlight this in a Proposition.

Proposition 3 *The NRQ property holds in Nash equilibrium under strict queue priority.*

Now $t_1 = \tau_1$ so that Proposition 1 applies and $t_1 = t_0 + N/\psi$. We shall briefly review the analysis of the bottleneck model for the case of general concave scheduling preferences.

By concavity of u, t_0 is the unique solution to the equation

$$u(0,t_0) = u(0,t_0 + N/\psi).$$

The utility function is given by $u(q(a), a + q(a))$. We omit below the arguments of $u(\cdot)$ to economise on notation. The first-order condition for Nash equilibrium is $\frac{\partial u}{\partial a} = u_1 \cdot q'(a) + u_2 \cdot [1 + q'(a)] = 0$, $a \in [t_0, t_1]$. Using (4) leads to the equilibrium arrival rate

$$\rho(a) = \psi \frac{u_1}{u_1 + u_2} > 0, \tag{5}$$

which is strictly positive on $[t_0, t_1]$ by Condition 1. (Condition 2 is not necessary here.)

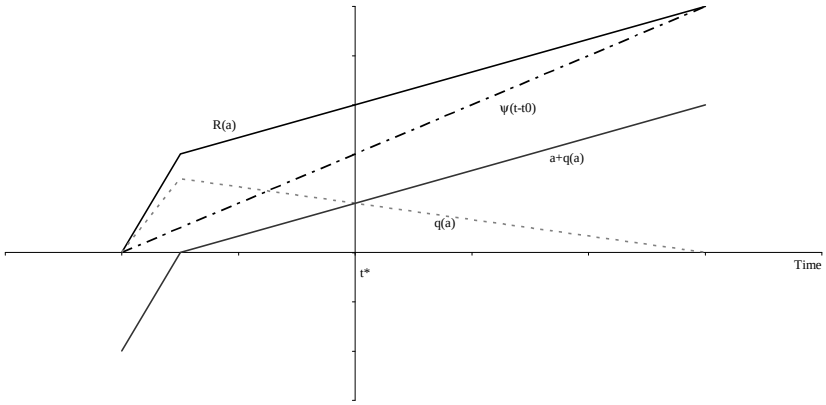

Figure 1: The evolution of the queue under strict queue priority with linear utility

By (5), $\rho(a) > \psi$ exactly when $u_2 > 0$, which occurs exactly when $a+q(a) < 0$. Thus the queue builds up until time $\tilde{a} < 0$ defined by $\tilde{a} + q(\tilde{a}) = 0$, at which time the queue begins to diminish.

The arrival rate is decreasing. To see this for $a \neq \tilde{a}$, differentiate the equilibrium condition twice to find

$$(q'(a), 1+q'(a)) \begin{pmatrix} u_{11} & u_{12} \\ u_{12} & u_{22} \end{pmatrix} (q'(a), 1+q'(a))^T + (u_1+u_2) q''(a) = 0.$$

The first term here is negative since $u(\cdot)$ is concave, and hence the second term is positive. Then $q''(a) \geq 0$ by Condition 1. Find from (4) that $\rho'(a)/\psi = q''(a)$, such that $\rho'(a) \geq 0$. The utility function is not required to be differentiable at the point $(q(\tilde{a}), \tilde{a} + q(\tilde{a}))$.

For any small $\varepsilon > 0$, we have $u_2(q(\tilde{a}+\varepsilon), \tilde{a}+\varepsilon+q(\tilde{a}+\varepsilon)) < 0$ and $0 < u_2(q(\tilde{a}-\varepsilon), \tilde{a}-\varepsilon+q(\tilde{a}-\varepsilon))$, while $u_1(q(a), a+q(a)) < 0$. Hence $\rho(\cdot)$ can only jump down at $\tilde{a}$. Such a jump occurs in the linear case, where the arrival rate is $\rho(a) = \psi\frac{\alpha}{\alpha-\beta}$ for $a < \tilde{a}$, and $\rho(a) = \psi\frac{\alpha}{\alpha+\gamma}$for $a > \tilde{a}$, which is piecewise constant with a downward jump at $\tilde{a} = -\frac{\beta}{\alpha}\frac{\gamma}{\beta+\gamma}\frac{N}{\psi}$.

Figure 1 shows the evolution of the queue under strict queue priority with linear utility. The curve $R(a)$ is the cumulative arrival rate, the kink occurs at the time where users exit at time $t^* = 0$. The curve $\psi(t-t_0)$ represents the cumulative number of exits from the queue. The curve $q(a)$ shows the duration in the queue for users entering the queue at time a. It is maximal for users who exit at time t^*. The curve $a+q(a)$ indicates the exit time for users entering the queue at time a.

4 No queue priority

With no queue priority, users to exit at any time are chosen at random at the rate ψ such that all users present in the queue have the same chance to exit. We first formalise this notion and show that if there is a residual queue at the time t_1 of the last arrival at the bottleneck, then the distribution of exit times conditional of being in the queue at time t_1 is uniform. Using this result, we then show that the acceptable lateness condition 2 is sufficient to guarantee the NRQ property in Nash equilibrium under no queue priority and that the equilibrium arrival rate is indeed positive. The acceptable lateness condition cannot be relaxed in general.

We formulate the no queue priority assumption by means of the hazard rate using concepts and results from duration analysis (Lancaster, 1990). The hazard rate does not depend on a as all users present in the queue at time t have the same probability to exit. Define the hazard rate of a user who is present in the queue at time t as

$$\lambda(t) = \frac{f(t|a)}{1 - F(t|a)} = \frac{\psi}{Q(t)}, \tag{6}$$

where $f(t|a)$ and $F(t|a)$ are respectively the density and cumulative distribution of exit time t conditional on being in the queue at time a. The survivor function $1 - F(t|a)$ can be expressed in terms of the integrated hazard by

$$1 - F(t|a) = e^{-\int_a^t \lambda(s)ds}. \tag{7}$$

The following technical Lemma concerns the conditional density of exit times when there is a residual queue after the last arrival. It states that when a pool of users exit with equal probability at a constant rate during some interval, then the exit time for each of them is uniformly distributed over this interval.

Lemma 2 *Consider the no queue priority case. Let t_1 be the time of the last arrival and assume that $Q(t_1) > 0$. Then the exit time conditional on being in the queue at time a $(t_1 \leq a \leq \tau_1)$ is uniformly distributed over the interval $[a, \tau_1]$ with $f(t|a) = \lambda(a), t \in [a, \tau_1]$. Furthermore, $\lambda'(a) = \lambda^2(a)$.*

We shall now show that concave utility as defined above together with the acceptable lateness condition 2 is sufficient to establish the no residual queue property for the no queue priority case. The acceptable lateness condition states that the marginal disutility of lateness is smaller than the marginal disutility of duration in the queue. If the queue diminishes quickly enough as arrival time increases, users will then postpone arrival until the queue is no longer decreasing so quickly. The second half of the Proposition establishes that condition 2 is also necessary for the NRQ property under linear utility. Hence condition 2 cannot be relaxed in general.

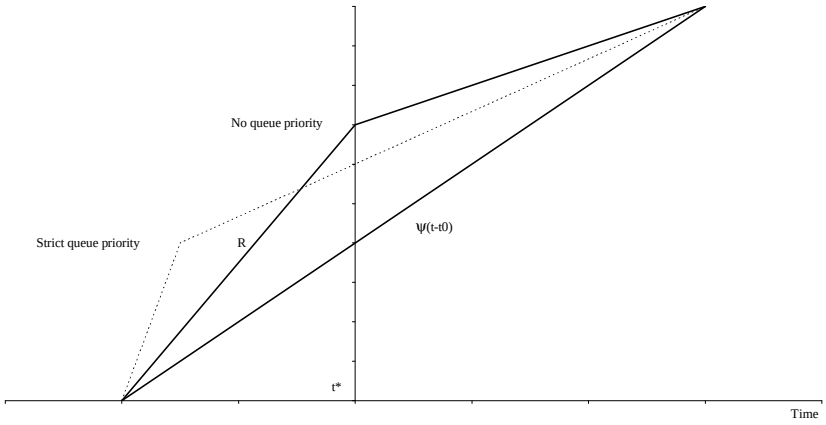

Figure 2: The evolution of the queue under no queue priority with linear utility

Proposition 4 *The acceptable lateness condition 2 is sufficient for the no residual queue property to hold. Under linear utility, condition 2 is also necessary.*

Proposition 5 establishes that the equilibrium arrival rate is always strictly positive under the acceptable lateness condition 2 and that the condition cannot be relaxed in general.

Proposition 5 *The acceptable lateness condition 2 is sufficient for the equilibrium arrival rate to be strictly positive over the interval* $[t_0, t_1]$ *defined by* $u(0, t_0) = u(0, t_1)$. *Under linear utility, condition 2 is also necessary.*

Figure 2 illustrates the evolution of the queue under no queue priority and linear utility. For comparison, the figure also shows the evolution of the queue under strict queue priority. The kinked curves are the cumulative arrival rates. Note that in the NQP case, the kink in the cumulative arrival rate occurs at time $t^* = 0$. The straight curve represents the cumulative number of exits from the queue.

5 Loose queue priority

This section concerns the case of loose queue priority, which we shall define as an intermediate case between the cases examined so far of strict and no queue priority. We shall show that the acceptable lateness condition 2 is sufficient to establish the no residual queue property for the case of loose queue priority; hence Condition 2 implies that Proposition 1 holds.

Under strict queue priority, users exit strictly in the order in which they arrive. Under no queue priority, users present in the queue at any instant all have the same

probability of exit. The intermediate case of loose queue priority is defined by requiring that at any instant, users whose present duration in the queue is longer have a higher chance to exit than users whose present duration in the queue is shorter. So arrival time matters, even if queue priority is not strict. There are very many possibilities for explicitly defining processes that have this property. The example below provides one simple way to model loose priority.

Example 1 *Introduce a variable $N(a,t)$ denoting the number of users in the queue at time t who arrived at the queue after time $a, a \leq t$. We have $N(a,t) \leq Q(t)$. Furthermore, $N(t,t) = 0$ and $N(t_0,t) = Q(t)$. At time t, there are $Q(t) - N(a,t)$ users in the queue who arrived earlier than a. Users exit the queue at the rate ψ, but under loose queue priority the hazard is not the same for everybody, it depends on the time of arrival a. We want the hazard rate, denoted $\lambda(t|a)$, to increase with the duration of the stay in the queue. One possible way of achieving this is by specifying the hazard rate to be*

$$\lambda(t|a) = H\left(\frac{N(a,t)}{Q(t)}\right)\frac{\psi}{Q(t)},$$

where $H(\cdot)$ is an increasing density on the unit interval with $H(0) < 1$. This hazard rate increases with the duration in the queue. The definition encompasses strict and no queue priority as limiting cases as $H(\cdot)$ approaches either a point mass at 1 or a uniform density. The hazard for the last user has $\lambda(t|t_1) = H\left(\frac{N(t_1,t)}{Q(t)}\right)\frac{\psi}{Q(t)} = H(0)\frac{\psi}{Q(t)} < \frac{\psi}{Q(t)}$ $(t_1 \leq t)$.

Recall that t_1 is the time of the last arrival at the queue, while $\tau_1 = t_1 + Q(t_1)/\psi$ is the time of the last exit from the queue. When there is a residual queue $Q(t_1) > 0$ then $\tau_1 > t_1$.

In the case of no queue priority we noted in Proposition 4 that the acceptable lateness condition 2 implies that $Q(t_1) > 0 \Rightarrow E(u|\tau_1) > E(u|t_1)$, contradicting that we can have $Q(t_1) > 0$ in Nash equilibrium. In this case the distribution of exit times conditional on entry at time t_1 is the uniform distribution over the interval $[t_1, \tau_1]$. We denoted this by $F(t|t_1)$.

In the case of strict queue priority we noted that $Q(t_1) > 0 \Rightarrow u(\tau_1) > u(t_1)$, which again contradicts that we can have $Q(t_1) > 0$ in Nash equilibrium. This happens because the last user entering at time t_1 will exit at time τ_1 with probability 1.

In order to establish the no residual queue property for the case of loose priority, it is sufficient to give a condition on the distribution of exit times conditional on entry at time t_1. Denote this distribution by $\tilde{F}(\cdot|t_1)$. We require that loose queue priority satisfies the following condition.

Condition 3 *(Loose queue priority) Under loose queue priority, the distribution of exit times conditional on arriving last, $\tilde{F}(\cdot|t_1)$, first-order stochastically dominates $F(\cdot|t_1)$, where $F(\cdot|t_1)$ is the uniform distribution over $[t_1, \tau_1]$ with $\tau_1 = t_1 + Q(t_1)/\psi$.*

The loose queue priority condition immediately implies that if there is a residual queue, then the last user to arrive is worse off under loose queue priority than under no queue priority (the utility function is decreasing in exit time, for any given arrival time). Hence Proposition 4 leads naturally to the following Proposition.

Proposition 6 *Under loose queue priority, the acceptable lateness condition 2 implies the no residual queue property in Nash equilibrium.*

Hence Condition 2 is sufficient to ensure that Proposition 1 applies, also in the case of loose queue priority.

6 Concluding remarks

This paper has considered a generalised version of the Vickrey bottleneck model of congestion users having general concave utility defined over the duration in the queue as well as the time of exit from the queue. The queue may be subject to varying degrees of random sorting, ranging from strict FIFO queue priority to no queue priority. The no residual queue (NRQ) property holds when the queue has vanished at the time of the last arrival. Proposition 1 shows that the NRQ property is sufficient to derive a number of results that are useful for designing policies to regulate congestion. In particular, the interval of arrival as well as the expected utility of users are independent of the queueing regime, provided the NRQ property holds. The remainder of the paper then establishes that the acceptable lateness condition 2, restricting the relation between the marginal utilities of duration and exit time, is sufficient for the NRQ property to hold in Nash equilibrium under all queueing regimes considered and that this condition cannot be relaxed in general.

For simplicity, we have only considered the case where total usage is constant. The extension to endogenous total demand is however straightforward.

The paper leaves open the characterisation of Nash equilibrium when the NRQ property does not hold. In that case, the convenient results of Proposition 1 are not available. The paper also leaves open the question of what happens under random queue sorting when the acceptable lateness condition is not satisfied. It is possible that there are combinations of queueing regimes and strictly concave utility for which the NRQ property does hold.

We must acknowledge some further limitations of our analysis. A main simplification is that we assume homogenous users, whereas heterogeneity is likely in actual queueing situations. Lindsey (2004) presents an analysis of user heterogeneity for the bottleneck model with strict FIFO queue and scheduling utility which is separable in duration in the queue and time of exit from the queue. It may be possible to extend Lindsey's analysis to allow for random queue sorting. We leave this for the future.

References

Arnott, R. A., de Palma, A. and Lindsey, R. (1993) 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, A. and Lindsey, R. (1996) Information and usage of free-access congestible facilities with stochastic capacity and demand *International Economic Review* **37**(1), 181–203.

Arnott, R. A., de Palma, A. and Lindsey, R. (1999) Information and time-of-usage decisions in the bottleneck model with stochastic capacity and demand *European Economic Review* **43**(3), 525–548.

Arnott, R. J. (1979) Unpriced transport congestion *Journal of Economic Theory* **21**(2), 294–316.

Barro, R. J. and Romer, P. M. (1987) Ski-Lift Pricing, with Applications to Labor and Other Markets *American Economic Review* **77**(5), 875–890.

Becker, G. S. (1991) A Note on Restaurant Pricing and Other Examples of Social Influences on Price *Journal of Political Economy* **99**(5), 1109–1116.

Blanc, J. P. C. (2009) Bad luck when joining the shortest queue *European Journal of Operational Research* **195**(1), 167–173.

Daniel, J. I. (1995) Congestion Pricing and Capacity of Large Hub Airports: A Bottleneck Model with Stochastic Queues *Econometrica* **63**(2), 327–370.

de Palma, A. and Arnott, R. A. (1989) The temporal use of a telephone line *Information Economics and Policy* **4**(2), 155–174.

de Palma, A. and Fosgerau, M. (2011) Dynamic and static congestion models: a review *in* A. de Palma, R. Lindsey, E. Quinet and R. Vickerman (eds), *A Handbook of Transport Economics* Edward Elgar chapter 9.

Hassin, R. (1985) On the Optimality of First Come Last Served Queues *Econometrica* **53**(1), 201–202.

International Transport Forum (2007) *The Extent of and Outlook for Congestion. Briefing Note*.

Knudsen, N. C. (1972) Individual and Social Optimization in a Multiserver Queue with a General Cost-Benefit Structure *Econometrica* **40**(3), 515–528.

Lancaster, T. (1990) *The Econometric Analysis of Transition Data* Econometric Society Monographs Cambridge University Press New York.

Lindsey, R. (2004) Existence, Uniqueness, and Trip Cost Function Properties of User Equilibrium in the Bottleneck Model with Multiple User Classes *Transportation Science* **38**(3), 293–314.

Naor, P. (1969) The regulation of queue size by levying tolls *Econometrica* **37**(1), 15–24.

Sattinger, M. (2002) A Queuing Model of the Market for Access to Trading Partners* *International Economic Review* **43**(2), 533–547.

Texas Transportation Institute (2007) *The 2007 Urban Mobility Report, September*.

Vickrey, W. S. (1969) Congestion theory and transport investment *American Economic Review* **59**(2), 251–261.

Yoshida, Y. (2008) Commuter arrivals and optimal service in mass transit: Does queuing behavior at transit stops matter? *Regional Science and Urban Economics* **38**(3), 228–251.

A Proofs

Proof of lemma 1.

Proof. All N users can arrive and be served without queueing during an interval of length N/ψ, so $-\infty < -N/\psi \leq \tau_0, \tau_1 \leq N/\psi < \infty$. There must be arrivals before the queue can start, so $t_0 \leq \tau_0$. If $t_0 < \tau_0$, some users can benefit from postponing arrival so $t_0 = \tau_0$ in equilibrium. Similarly, $t_1 \leq \tau_1$, since otherwise some users could benefit from arriving earlier. In equilibrium, there is always queue during $]\tau_0, \tau_1[$ since otherwise users could benefit from moving into the gap in the queue. The arrival rate is locally bounded so not all users can arrive at time 0. The first arrival time occurs strictly before the preferred exit time 0, since otherwise it would be possible to arrive at time 0 and be served immediately. Similarly, the last arrival time occurs strictly after time 0. ■

Proof of Proposition 1.

Proof. The NRQ property implies that $t_1 = \tau_1$, which means that $Q(t_1) = 0$. Hence the durations in the queue are zero at times t_0 and t_1 so that $u(0, t_0) = u(0, t_1)$. By Lemma 1, the queue lasts from t_0 to t_1 such that $N = \psi(t_1 - t_0)$. Consequently, t_0 and t_1 are unique due to concavity of $u(\cdot)$ and $t_0 < 0 < t_1$. By the equilibrium condition, $E(u|a) = u(0, t_0)$ for all $a \in [t_0, t_1]$. Differentiating $N = \psi(t_1 - t_0)$ leads to $1 = \psi\left(\frac{\partial t_1}{\partial N} - \frac{\partial t_0}{\partial N}\right)$. Differentiating $u(0, t_0) = u(0, t_1)$ leads to $u_2(0, t_0)\frac{\partial t_0}{\partial N} = u_2(0, t_1)\frac{\partial t_1}{\partial N}$, so that

$$\frac{\partial t_0}{\partial N} = \frac{1}{\psi}\frac{u_2(0, t_1)}{u_2(0, t_0) - u_2(0, t_1)} < 0.$$

Then

$$\frac{\partial u(0, t_0)}{\partial N} = \frac{1}{\psi}\frac{u_2(0, t_0)\, u_2(0, t_1)}{u_2(0, t_0) - u_2(0, t_1)} < 0.$$

Straightforward computation establishes that when $u(\cdot)$ is concave, then the marginal utility decreases

$$\frac{\partial^2 u(0, t_0)}{\partial N^2} = \frac{1}{\psi^2}\frac{u_2(0, t_0)^3 u_{22}(0, t_1) - u_2(0, t_1)^3 u_{22}(0, t_0)}{(u_2(0, t_0) - u_2(0, t_1))^3} \leq 0,$$

with strict inequality when $u(\cdot)$ is strictly concave. ■

The following Lemma collects some relationships between the hazard rate and the corresponding conditional density and cumulative distribution function. We will use the results in the Lemma many times in the proofs below and will therefore omit references to the Lemma.

Lemma 3 *Let the hazard rate λ and the corresponding $f(t|a)$ and $F(t|a)$ be as defined above. Then the following relations hold.*

$$\begin{aligned} f(a|a) &= \lambda(a) && (8)\\ \frac{\partial F(t|a)}{\partial a} &= -\frac{\lambda(a)}{\lambda(t)} f(t|a) && (9)\\ \frac{\partial f(t|a)}{\partial a} &= \lambda(a) f(t|a) && (10) \end{aligned}$$

Proof. The first assertion follows from (6), since $F(a|a) = 0$. Differentiate (7) to find that

$$\frac{\partial F(t|a)}{\partial a} = -\lambda(a) e^{-\int_a^t \lambda(s)ds} = -\lambda(a)(1 - F(t|a)).$$

Then the second assertion follows by substitution from (6), while the third assertion follows by differentiation with respect to t. ■

Proof of Lemma 2.

Proof. Evaluate first $1 - F(t|a)$. Let $t_1 \leq a \leq t \leq \tau_1$. Then by (7)

$$1 - F(t|a) = \exp\left(-\int_a^t \frac{\psi}{Q(t_1) - \psi(s - t_1)} ds\right),$$

where we use that $Q(s) = Q(t_1) - \psi(s - t_1)$. Make the substitution $x = Q(t_1)/\psi - (s - t_1)$ to find that

$$\begin{aligned} 1 - F(t|a) &= \exp\left(\int_{Q(t_1)/\psi-(a-t_1)}^{Q(t_1)/\psi-(t-t_1)} \frac{1}{x} dx\right)\\ &= \frac{Q(t_1)/\psi - (t - t_1)}{Q(t_1)/\psi - (a - t_1)} = \frac{\lambda(a)}{\lambda(t)}. \end{aligned}$$

Use (6) to see that $f(t|a) = \lambda(a)$. As the density of exit times conditional on a is constant, the exit time is uniformly distributed. To verify the last statement of the Proposition, simply differentiate

$$\frac{\partial \lambda(a)}{\partial a} = -\frac{\psi Q'(a)}{Q^2(a)} = \frac{\psi^2}{Q^2(a)} = \lambda^2(a).$$

■

Proof of Proposition 4.

Proof. Assume a Nash equilibrium with a residual queue at time t_1 and consider $a > t_1$. The expected utility at time a, given by (2), is

$$E(u|a) = \lambda(a) \int_a^{\tau_1} u(t-a,t)\, dt$$

by Lemma 2. Using the last statement of Lemma 2, the derivative with respect to the arrival time a is seen to be

$$\frac{1}{\lambda(a)} \frac{\partial E(u|a)}{\partial a} = E(u|a) - u(0,a) - \int_a^{\tau_1} u_1(t-a,t)\, dt. \tag{11}$$

Considering the following identity

$$u(\tau_1 - a, \tau_1) - u(0,a) = \int_a^{\tau_1} [u_1(t-a,t) + u_2(t-a,t)]\, dt,$$

we may write

$$\frac{1}{\lambda(a)} \frac{\partial E(u|a)}{\partial a} = E(u|a) - u(\tau_1 - a, \tau_1) + \int_a^{\tau_1} u_2(t-a,t)\, dt.$$

Add the two expressions for $\frac{\partial E(u|a)}{\partial a}$ to obtain

$$\begin{aligned} \frac{1}{\lambda(a)} \frac{\partial E(u|a)}{\partial a} &= \left[E(u|a) - \frac{1}{2}(u(0,a) + u(\tau_1 - a, \tau_1)) \right] \\ &+ \frac{1}{2} \int_a^{\tau_1} [u_2(t-a,t) - u_1(t-a,t)]\, dt \end{aligned}$$

The first term on the RHS is positive by Jensen's inequality since $u(t-a,t)$ is concave as a function of t and the second term is strictly positive by Condition 2. Thus, $E(u|a)$ is strictly increasing on $]t_1, \tau_1[$ so that

$$E(u|t_1) < E(u|\tau_1) = u(0,\tau_1), \tag{12}$$

which contradicts Nash equilibrium.

To verify the second assertion of the Proposition, note that in the linear case,

$$\begin{aligned} \frac{1}{\lambda(a)} \frac{\partial E(u|a)}{\partial a} &= \frac{1}{2} \int_a^{\tau_1} [u_2(t-a,t) - u_1(t-a,t)]\, dt \\ &= \frac{1}{2}(\tau_1 - a)(\alpha - \gamma). \end{aligned}$$

Then $\frac{\partial E(u|a)}{\partial a} > 0$ is equivalent to Condition 2 and so Condition 2 is also necessary. ■

Proof of Proposition 5.

Proof. The expression for the expected utility conditional on arrival at time a is (2). Using (10), we express the equilibrium condition for the no queue priority case as follows.

$$\frac{\partial E\left(u|a\right)}{\partial a} = \lambda\left(a\right) E\left(u|a\right) - u\left(0,a\right)\lambda\left(a\right) - E\left(u_1|a\right) = 0,$$

which can be solved using $\lambda\left(a\right) = \psi/Q\left(a\right)$ to yield

$$\frac{Q\left(a\right)}{\psi} = \frac{E\left(u|a\right) - u\left(0,a\right)}{E\left(u_1|a\right)}.$$

Differentiate again and use that (1) gives $Q'\left(a\right) = \rho\left(a\right) - \psi$ to find

$$\frac{\rho\left(a\right)}{\psi} = 1 - \frac{u_2\left(0,a\right)}{E\left(u_1|a\right)} - \frac{\frac{\partial E(u_1|a)}{\partial a}}{\lambda\left(a\right) E\left(u_1|a\right)}. \tag{13}$$

Multiply all terms in (13) by $-\lambda\left(a\right) E\left(u_1|a\right) > 0$ to find that $\rho\left(a\right) > 0$ iff

$$-\lambda\left(a\right) E\left(u_1|a\right) + \lambda\left(a\right) u_2\left(0,a\right) + \frac{\partial E\left(u_1|a\right)}{\partial a} > 0. \tag{14}$$

Carry out the differentiation using Lemma 3 to find that

$$\frac{\partial E\left(u_1|a\right)}{\partial a} = -\lambda\left(a\right) u_1\left(0,a\right) - E\left(u_{11}|a\right) + \lambda\left(a\right) E\left(u_1|a\right).$$

Insert this into the inequality (14) to find that it is equivalent to

$$\lambda\left(a\right)\left[u_2\left(0,a\right) - u_1\left(0,a\right)\right] - E\left(u_{11}|a\right) > 0. \tag{15}$$

The second term is positive since u is concave. Therefore Condition 2 implies that $\rho\left(a\right) > 0$.

When utility is linear, (13) shows that the equilibrium arrival rate is

$$\rho(a) = \begin{cases} \psi\frac{\alpha+\beta}{\alpha}, a < 0 \\ \psi\frac{\alpha-\gamma}{\alpha}, a > 0. \end{cases}$$

Then $\rho\left(a\right) > 0$ implies Condition 2. ■

Proof of Proposition 6.

Proof. Assume that $Q(t_1) > 0$. Then $E_{\tilde{F}}(u|t_1) \leq E_F(u|t_1)$, due to first-order stochastic dominance. But $E_F(u|t_1) < u(0, \tau_1)$ by (12) in the proof of Proposition 4. Then $E_{\tilde{F}}(u|t_1) < u(0, \tau_1)$ and the last user would prefer to arrive at τ_1 rather than at t_1. This contradicts Nash equilibrium. Hence we must have $Q(t_1) = 0$ in Nash equilibrium. ■