Random queues and risk averse users

André de Palma, Mogens Fosgerau

Random queues and risk averse users*

André de Palma† Mogens Fosgerau‡

December 16, 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 define 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 à Paris Dynamique de l'acceptabilité and PREDIT: scheduling, trip timing and scheduling preferences. We are grateful to the editor Lorenzo Peccati and three anonymous reviewers for their constructive comments. We also wish to thank 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 us a number of very useful comments.

†École Normale Supérieure de Cachan, Centre d'Economie de la Sorbonne and École Polytechnique, andre.depalma@ens.cachan.fr.

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

1 Introduction

We generalize 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. However, non-FIFO queuing is important in reduced-form models of all non-trivial networks, since person B who entered the network later may affect the level of performance received by person A who entered the network earlier. Downtown traffic congestion is one example of this; a swimming pool is another; and a telecommunication network is yet another. 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). An extreme case is a pure random queue[1], and 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. 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]

[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

The economic literature has previously paid attention to the properties of user equilibrium in queues with strict queue priority using the seminal Vickrey (1969) bottleneck model. This model offers many insights that are central to the understanding of congested demand peaks. Arnott, de Palma and Lindsey (1993) summarize 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.[3] The bottleneck serves users at a fixed rate. A continuum of users choose their time of arrival at bottleneck into the queue to minimize 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.[4]

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

property.

[3]The operations research literature considers systems of bottlenecks but with exogenous arrival rate (e.g. Osorio and Bierlaire, 2009).

[4]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). Glazer and Hassin (1983) discusses endogenisation of the arrival rate in a ?/M/1 queue, finding with a compact service interval that the arrival rate is not constant.

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 organized as follows. Section 2 presents the general framework, 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 generalizes the standard case of *strict queue priority* and establishes that the NRQ property holds here. Next, Section 4 considers the opposite 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 rate ρ will be determined endogenously within the model as a consequence of individual decisions. 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[5]

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

[5] $1_{\{\cdot\}}$ is the indicator function for the event in curly brackets.

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 $Q(t_1) > 0$.

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 formalize 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 note that this means, in particular, that all users have the same preferred exit time t^*. 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 $t > t^*$. The common preferred exit time is set to zero, $t^* = 0$, without loss of generality.

Users choose their arrival time a to maximize 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* $t > 0$.

Note that we do not impose a condition on derivatives at $t = 0$. We have not required that utility is differentiable at these points, which allows utility to have a kink as is the case when utility is linear, which is the case investigated by Vickrey (1969) and Arnott et al. (1993). The case of linear utility will be important for results and also helps in facilitating interpretation of results. The linear utility formulation is[6]

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

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) summarizes 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.[7] 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 a continuum of homogenous users, this requirement turns into the condition that the expected utility is constant over the times at which users arrive, i.e. over the support of ρ, and not strictly larger at any other time.

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

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

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

to which optimality is defined. Denote by $\lambda(a)$ a time varying toll depending on the arrival time at the bottleneck. We define utility to be money-metric utility with any toll payment being simply subtracted, such that utility is $u(d,t) - \lambda(a)$. 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 $\rho(a) > 0$. This leads to the equation

$$-u(0,a) f(a|a) + \int_a^\infty \left[u(t-a,t) \frac{\partial f(t|a)}{\partial a} - u_1(t-a,t) f(t|a) \right] dt = 0.$$

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 *In Nash equilibrium, the support of Q is a finite interval 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 $u(0,t_0) = u(0,t_1)$. 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 equilibria where the NRQ property holds. Then for any N, 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 $u(0,t_0) = u\left(0, t_0 + \frac{N}{\psi}\right)$. The expected utility of any user is $u(0,t_0)$. The marginal change in expected utility from additional users, at Nash equilibria, is*

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

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 in Nash equilibrium under strict, loose and no queue priority and hence that Proposition 1 applies in all these regimes.

The following Proposition summarizes some properties of Nash equilibrium under a toll which eliminates queueing. The Proposition considers the optimal time varying toll, which eliminates queueing, meaning that it leads a unique Nash equilibrium with $Q(t) = 0$ for all t and $\rho(t) = \psi$ at all times t in the arrival interval $[t_0, t_1]$. Since there is never any queue under this toll, then the Nash equilibrium is not affected by random queue sorting.

Proposition 2 *Let $\hat{t}_0$ be the first arrival time in Nash equilibrium with the NRQ property as given in Proposition 1. Impose a time varying toll that depends on the arrival time a at the bottleneck given by*

$$\lambda(a) = \left[u(0,a) - u\left(0,\hat{t}_0\right)\right]^+.$$

Then there exists a unique Nash equilibrium; it has $t_0 = \hat{t}_0$, departure rate $\rho = \psi$ during $[t_0, t_1]$, and $Q(t) = 0$ for all t.

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 unique existence of Nash equilibrium is easy to establish. Utility is constant during $[t_0, t_1]$ in Nash equilibrium, such that

$$q'(a) = -\frac{u_2(q(a), a + q(a))}{u_1(q(a), a + q(a))}. \tag{5}$$

With $[t_0, t_1]$ determined as in Proposition 1, the first-order differential equation (5) together with (4) determines $\rho(\cdot)$. It is immediate that this arrival rate supports a Nash equilibrium and hence we have existence. Is is also immediate that any Nash equilibrium must satisfy the same equations and hence we have uniqueness. 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 *Under strict queue priority, there exists a unique Nash equilibrium and it exhibits the NRQ property.*

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 economize 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

$$\rho(a) = \psi \frac{u_1}{u_1 + u_2} > 0, \tag{6}$$

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

By (6), $\rho(a) > \psi$ 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 $0 <$

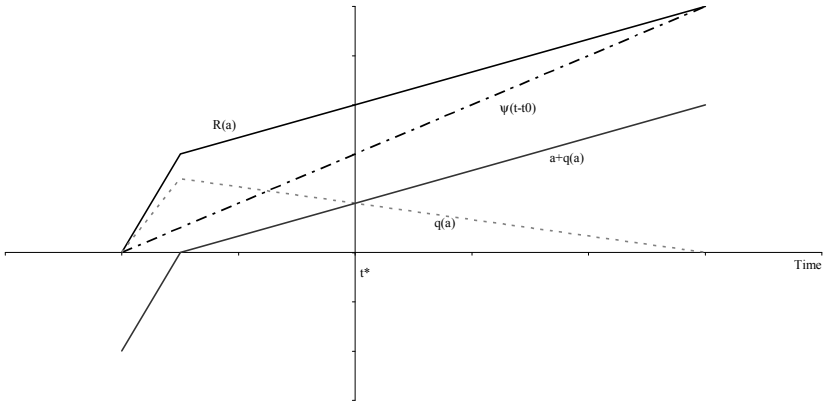

Figure 1: The evolution of the queue under strict queue priority with linear utility

$u_2\left(q\left(\tilde{a}-\varepsilon\right), \tilde{a}-\varepsilon+q\left(\tilde{a}-\varepsilon\right)\right)$, while $u_1\left(q\left(a\right), a+q\left(a\right)\right)<0$. Hence $\rho\left(\cdot\right)$ 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 $R\left(a\right)$ is the cumulative arrival rate, the kink occurs at the time where users exit at time $t^*=0$. The curve $\psi\left(t-t_0\right)$ represents the cumulative number of exits from the queue. The curve $q\left(a\right)$ 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 $a+q\left(a\right)$ 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 formalize 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

$$\lambda(t) = \frac{f(t|a)}{1 - F(t|a)} = \frac{\psi}{Q(t)}, \tag{7}$$

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

$$1 - F(t|a) = e^{-\int_a^t \lambda(s)ds}. \tag{8}$$

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.

Proposition 4 *Under no queue priority, 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 *Under no queue priority, 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.*

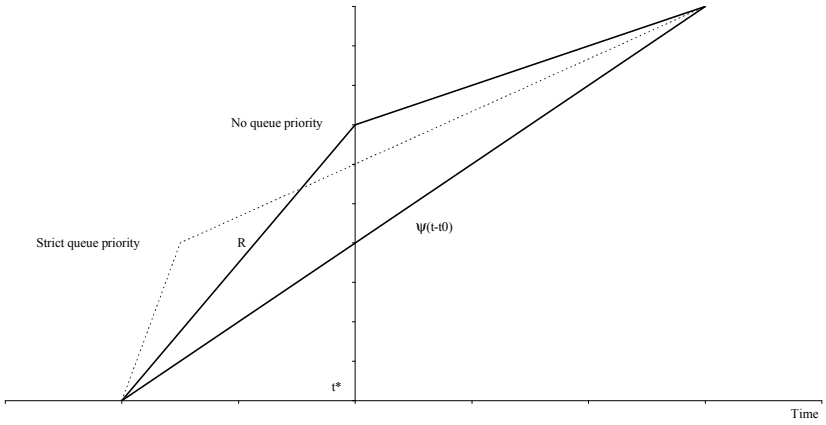

Figure 2: The evolution of the queue under no queue priority with linear utility

The proof of Proposition 5 does not guarantee existence of Nash equilibrium. However, under linear utility, the proof establishes a first-order differential equation for the equilibrium arrival rate. As in the case of strict queue priority, this both gives us existence, since this arrival rate supports Nash equilibrium, and uniqueness, since any Nash equilibrium must satisfy the conditions that were used to construct the arrival rate.

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. As in the case of no queue priority, we cannot guarantee existence of Nash equilibrium. However, 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 *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 generalized 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. It should, however, be acknowledged that this condition is also quite restrictive. Under linear utility it rules out that a minute of lateness could be more costly than a minute of travel time. In particular it rules out that there could be a large penalty that effectively ruled out lateness.

The NRQ property is not universal. It is easy to construct straightforward cases where it does not hold. An example is strict queue priority with linear utility but with infinite cost of lateness, such that late exit is ruled out. In this case, there is a unique Nash equilibrium in which users arrive at a constant rate greater than capacity; arrivals stop at some time $t_1 < t^*$ and there is a strictly positive queue at this time; the queue dissipates during $[t_1, t^*]$ and has vanished exactly at time t^*.

Proposition 1 crystallizes the insight that the NRQ property allows the equilibrium utility to be determined just as a function of the number of users. This insight

is behind the analyses of optimal time varying tolls in the bottleneck model, step tolls, as well as fast lane mechanisms. Common to these analyses is that they consider ways of manipulating the queue, for example by extracting revenue, that do not upset the NRQ property which ensures that the equilibrium utility is unaffected. Users are neutral with respect to any policy that does not affect the NRQ property. This paper has made explicit that the NRQ property underlies these insights and shown that it is robust within some limits to random queue sorting. It is then straightforward that, e.g., step tolling and fast laning will have similar consequences for users under random queue sorting as under strict queue priority, although they may entail different consequences for toll revenues.

For simplicity, we have only considered the case where total usage N is constant. The extension to endogenous total demand is however straightforward. A way to proceed is be to let aggregate demand depend on the average utility obtained in equilibrium; let demand be strictly increasing as a function of average equilibrium utility and assume that demand tends to 0 as average equilibrium utility tends to minus infinity. Then note that Proposition 1 states that the average equilibrium is strictly decreasing as function of the number of users. Conditional on the unique existence of Nash equilibrium for each value of N, this is sufficient to guarantee that Nash equilibrium exists uniquely when N is allowed to be endogenous. All results in the paper then generalize to the case of endogenous demand.

The paper leaves open the characterization 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, but we have not been able to do it. The simplest way to go would be to repeat an analysis of heterogeneity that can be carried out for the case of strict queue priority and linear utility. In this case, let t^*, the preferred exit time, be heterogeneous. Then there would be only two distinct departure rates, one for users exiting early and one for users exiting late. Users would sort on t^* such that users arriving early would also arrive during the period where the early departure rate prevails and users arriving late would also arrive during the period where the late arrival rate prevails. There would be a definite time at which the switch from early to late

occurred. This property is central to the analysis of heterogeneity in t^* under strict queue priority and linear utility. However, this property breaks down under random queue sorting; then there would be a range where randomness meant it would not be possible to know whether a user would be early or late. We therefore leave the issue of heterogeneity for future research.

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.

Glazer, A. and Hassin, R. (1983) ?/M/1: On the equilibrium distribution of customer arrivals *European Journal of Operational Research* **13**(2), 146–150.

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.

Osorio, C. and Bierlaire, M. (2009) An analytic finite capacity queueing network model capturing the propagation of congestion and blocking *European Journal of Operational Research* **196**(3), 996–1007.

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. ■

Proof of Proposition 2.

Proof. The departure rate stated in the proposition supports a Nash equilibrium, since utility including the toll is constant and equal to $u(0, t_0)$ for users during

$[t_0, t_1]$ and strictly lower outside. Any other departure rate with first arrival at $\hat{t}_0$ would either lead to queueing or to exit later than t_1, which would lead to some users achieving lower utility. Any other first arrival time than t_0 would lead to utility lower than $u(0, t_0)$, either for the first or the last user. ■

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.*

$$f(a|a) = \lambda(a) \tag{9}$$

$$\frac{\partial F(t|a)}{\partial a} = -\frac{\lambda(a)}{\lambda(t)} f(t|a) \tag{10}$$

$$\frac{\partial f(t|a)}{\partial a} = \lambda(a) f(t|a) \tag{11}$$

Proof. The first assertion follows from (7), since $F(a|a) = 0$. Differentiate (8) 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 (7), 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 (8)

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

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 (7) 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

$$\frac{1}{\lambda(a)} \frac{\partial E(u|a)}{\partial a} = E(u|a) - u(0, a) - \int_a^{\tau_1} u_1(t-a, t)\, dt. \tag{12}$$

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

$$E(u|t_1) < E(u|\tau_1) = u(0, \tau_1), \tag{13}$$

which contradicts Nash equilibrium.

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

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

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 (11), we express the equilibrium condition for the no queue priority case as follows.

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

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

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

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

$$\frac{\rho(a)}{\psi} = 1 - \frac{u_2(0,a)}{E(u_1|a)} - \frac{\frac{\partial E(u_1|a)}{\partial a}}{\lambda(a)E(u_1|a)}. \tag{14}$$

Multiply all terms in (14) by $-\lambda(a)E(u_1|a) > 0$ to find that $\rho(a) > 0$ iff

$$-\lambda(a)E(u_1|a) + \lambda(a)u_2(0,a) + \frac{\partial E(u_1|a)}{\partial a} > 0. \tag{15}$$

Carry out the differentiation using Lemma 3 to find that

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

Insert this into the inequality (15) to find that it is equivalent to

$$\lambda(a)\left[u_2(0,a)-u_1(0,a)\right] - E(u_{11}|a) > 0. \tag{16}$$

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

When utility is linear, (14) 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 $\rho(a) > 0$ 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 (13) 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. ■