Random queues and risk averse users*

André de Palma† Mogens Fosgerau‡

August 28, 2009

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

PRELIMINARY

*We are grateful to Robin Lindsey, Katrine Hjorth, Hugo Harari-Kermadec, Søren Feodor Nielsen, Ken Small and seminar participants at the Swedish Royal Institute of Technology for comments. Mogens Fosgerau is supported by the Danish Social Science Research Council.

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

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

1 Introduction

Queueing phenomena are ubiquitous. There are queues in supermarkets, banks, public offices, restaurants [4], movie theatres, concert ticket sales, at ski lifts [3] and toll road booths, in airports [6], computer systems, communications systems, web services, call centers, and numerous other systems. Especially congestion during the commuting journey has been studied extensively [see 16]. Enormous amounts of time are lost queueing. Just for private transportation, the cost of congestion in Europe and the US has been estimated to be equivalent to more than 1 percent of GDP [10, 17].

Queues can be served under different queueing regimes [14, 8]. A useful benchmark is the first-in-first-out (FIFO) queue discipline, where users are served in the sequence of their arrival to the queue. Many real queues, however, involve an element of random sorting, so that queue priority is only partly maintained. For example, while FIFO applies to individual checkout lines in the supermarket, FIFO does not apply to the supermarket checkout system as a whole, when there is more than one line, since one user could arrive at the checkout later than another user and still be served before him [5]. Queues can also have random possibilities for overtaking; for example when a new checkout line opens in the supermarket.

A polar case to the FIFO queue is the pure random queue, i.e. a queue with no priority.[1] In this case, every person present in the queue at a given time has the same probability of being served as any other person in the queue. This corresponds, e.g., to a (virtual) queue to get through on a busy telephone line. Having been trying to place a call for a longer time does not increase the likelihood of being able to have access to the line [7].

The general practice in the operations research literature is to consider the arrival rate as exogenous, perhaps allowing the user to balk when meeting a long queue [13, 11]. In contrast, we consider a framework where the arrival rate into the queue is endogenous. In other words, the time of use is a decision variable of the user. For simplicity, we consider only the case where total usage is constant – the extension to endogenous total demand is straightforward.

The economic literature has previously paid attention to the properties of user equilibrium in queues with strict queue priority using the deterministic bottleneck model of Vickrey [18]. In his seminal framework, the arrival pattern depends on the evolution of the queue. Individuals are assumed to choose their time of arrival into the queue to minimise a weighted sum of the time spent in the queue, and linear penalties associated with being early or late at the destination. We extend Vickrey's model in two ways: by allowing for more general preference structures

[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 [9].

and by allowing for stochastic rather than deterministic service time. Individual risk aversion matters when service time is stochastic. Assuming a continuum of users, we are able to characterise the equilibrium arrival rate into the queue as well as the equilibrium user utility and the equilibrium time interval of arrivals.

The classic Vickrey bottleneck model offers many insights that are central to the understanding of congested demand peaks. Arnott et al. [1] summarise a number of these. The central point of the model is the description of the congestion cost as an externality related to the scheduling decisions of users and in particular that, in aggregate and in equilibrium, the cost associated to scheduling is as large as the average cost associated to the time spent queueing. This paper extends these insights to queues subject to random sorting, considering users with general concave utility of duration in the queue as well as of earliness/lateness.

The paper is organised as follows. Section 2 presents the general framework and then introduces the no residual queue (NRQ) property for a queue with a general random sorting mechanism. The NRQ property holds when the queue has exactly vanished at the time of the last arrival. We show that the NRQ property is sufficient to establish a number of useful results. In particular, we derive the equilibrium utility and the marginal utility of adding users under Nash equilibrium. Remarkably, these results are not affected by the random sorting mechanism, provided the NRQ property holds. Thus, the NRQ property is extremely useful in the analysis of congested demand peaks.[2]

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 polar case of no queue priority (users to be served are chosen completely at random from the queue). We establish also the NRQ property for this case. It turns out that in order for the NRQ property to hold in general, it is necessary to assume that the marginal disutility of being late is always less than the marginal disutility of duration in the queue. This means that users must be always willing to arrive one minute later in exchange for spending one minute less in the queue.

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 sufficient to guarantee the NRQ property to hold in general when queue priority is loose. Some concluding remarks are provided in Section 6.

[2]The NRQ property does not hold if capacity is stochastic [2]. We are thus restricting attention to cases where capacity can be regarded as fixed.

2 Model specification

2.1 Evolution of the queue

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[3]

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

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 $Q(t_1) > 0$.

2.2 Queue discipline

We shall consider various queue discipline regimes, i.e. ways of determining which user exits the queue at each instant. At one extreme we have the *strict queue priority* case, which was considered in the Vickrey [18] bottleneck model. Here the queue obeys the first-in-first-out principle (FIFO) and users exit strictly in the sequence in which they arrive. 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 a user to exit from the queue does not depend on how much time he has spent in the queue. It is the same for all users present 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.

[3] Denote by $1_{\{\cdot\}}$ the indicator function for the event in curly brackets.

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 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.

2.3 User preferences

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. Write his utility as $u(d, t - t^*)$. We take utility to be concave, to have a unique maximum at $d = 0$ for any $t - t^*$ and to have 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 consider users with identical t^* and take $t^* = 0$ 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. Clearly, users who exit early always satisfy this condition. This second condition is assumed where indicated.

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

We shall refer to the special case of linear utility. This will be important for results and also helps in facilitating interpretation of results. The linear utility formulation is[4]

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

where the parameters α, β and γ are strictly positive. For the linear case, Condition 1 states that $\beta < \alpha$, while Condition 2 states that $\gamma < \alpha$.

2.4 Nash equilibrium

We consider Nash equilibrium in pure strategies as the benchmark for rational behavior.[5] 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 minimal over the times at which users arrive, i.e. over the support of ρ.

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 $\rho(a) > 0$, i.e.,

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

Recall that t_0 and t_1 are the times of the first and the last arrival. The following Lemma provides some properties of the Nash equilibrium.

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

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

[4] The function x^+ equals $\max(x, 0)$, $x^- = (-x)^+$, so $x = x^+ - x^-$.

[5] The equilibrium concept is discussed by Arnott et al. [1].

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. ■

An implication of Lemma 1 is that $\rho(t_0) > \psi$, since the queue begins at time t_0. The no residual queue property states that there is no queue at the time of arrival of the last user.

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

This is an important property as it ensures that $t_1 = \tau_1$ in Nash equilibrium by Lemma 1. A main result of this paper is that the NRQ property holds in Nash equilibrium under strict, loose and no queue priority. The NRQ property allows us to further characterise the Nash equilibrium.

Proposition 1 *Consider Nash equilibrium in a queue where the Nash equilibrium is such that the no residual queue 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* $u(0, t_0) = u\left(0, t_0 + \frac{N}{\psi}\right)$*. The expected utility of any user is* $u(0, t_0)$*. The marginal change in expected utility from additional users 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.

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 mar-

ginal 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 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. These results are independent of the queueing regime. The following corollaries follow immediately.

Corollary 1 *In the linear case,* $\frac{\partial E(u|a)}{\partial N} = -\frac{1}{\psi}\frac{\beta\gamma}{\beta+\gamma}$ *and* $\frac{\partial^2 E(u|a)}{\partial N^2} = 0$.

Corollary 2 *The expected utility and the marginal expected utility (3) in Nash equilibrium in a queue with the no residual queue property are not affected by random queue sorting.*

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.

3 Strict queue priority

This is the case considered by Vickrey [18] and Arnott et al. [1] in the context of transportation and telecommunication. Here we consider a more general formulation of user preferences than the one considered by these authors. 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 no residual queue 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. Hence $t_1 = \tau_1$ so that Proposition 1 applies and $t_1 = t_0 + N/\psi$. 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.)

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 from the queue at the rate ψ, i.e., all users present in the queue at a given instant have the same chance to exit. The next Section 4.1 formalises this notion and derives the distribution of exit times after the time of the last arrival t_1 assuming that there is a residual queue at time t_1. Using this distribution, Section 4.2 establishes that the no residual queue property necessarily holds in Nash equilibrium in the no queue priority case, while Section 4.3 shows that the equilibrium arrival rate is positive. Condition 2 is sufficient to guarantee these results for general risk averse users

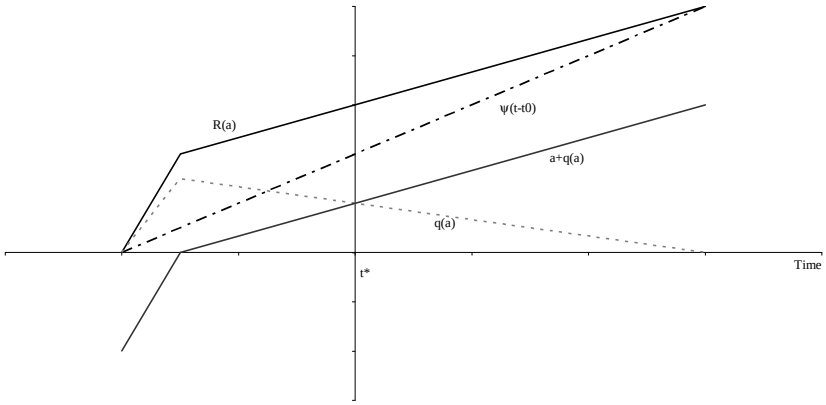

Figure 1: The evolution of the queue under strict queue priority with linear utility

consistent with our demand setting.

4.1 The hazard rate

We formulate the no queue priority assumption by means of the hazard rate using concepts and results from duration analysis [12]. 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 Lemma collects some relationships between the hazard rate and the corresponding conditional density and cumulative distribution function.

Lemma 2 *Let the hazard rate λ and the corresponding $f(t|a)$ and $F(t|a)$ be as*

defined above. Then the following relations hold.

$$f(a|a) = \lambda(a) \tag{8}$$

$$\frac{\partial F(t|a)}{\partial a} = -\frac{\lambda(a)}{\lambda(t)} f(t|a) \tag{9}$$

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

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. ■

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 3 *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, $\lambda'(a) = \lambda^2(a)$.*

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

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

■

4.2 The no residual queue property

We shall now show that concave utility as defined above together with Condition 2 is sufficient to establish the no residual queue property for the no queue priority case. Condition 2 ensures that the marginal disutility of lateness is smaller than the marginal disutility of duration in the queue, so that users will always prefer arriving one minute later to staying one minute longer 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.

Proposition 2 *Condition 2 is sufficient for the no residual queue property to hold. Under linear utility, Condition 2 is also necessary.*

Proof. Assume a Nash equilibrium with a residual queue at time $a \geq t_1$. We consider $a > t_1$ and note that the result for $a = t_1$ follows by taking limits. 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 3. Using the last statement of Lemma 3, 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

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

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,

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

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

Note that we have not ruled out the existence of Nash equilibrium when Condition 2 does not hold, but in that case the NRQ property does not hold in general. Evidence concerning preferences for the scheduling of commuting trips seems to indicate $\alpha < \gamma$ in that case [15].

4.3 The equilibrium arrival rate

This section investigates the equilibrium arrival rate under no queue priority. Proposition 3 establishes that the equilibrium arrival rate is always positive under Condition 2.

Proposition 3 *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.*

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

$$\frac{\rho\left(a\right)}{\psi}=1-\frac{u_{2}\left(0,a\right)}{E\left(u_{1}|a\right)}-\frac{\frac{\partial E\left(u_{1}|a\right)}{\partial a}}{\lambda\left(a\right)E\left(u_{1}|a\right)}. \tag{13}$$

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 2 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. ■

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.

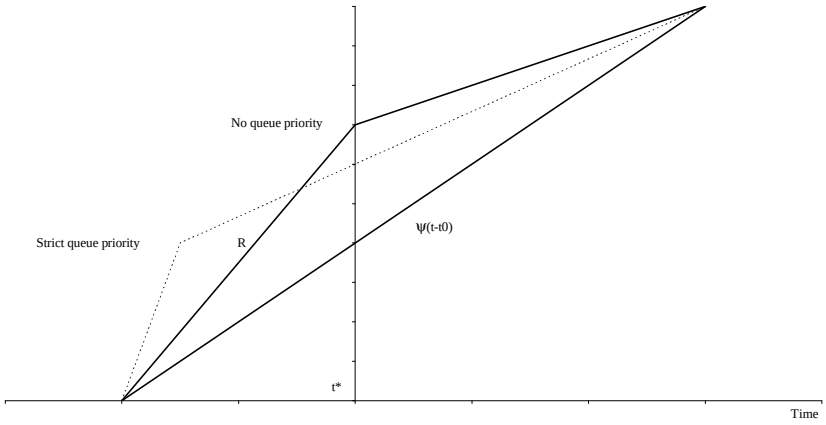

Figure 2: The evolution of the queue under strict queue priority with linear utility

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 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.

5.1 Motivation and illustration

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 a 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* $\lambda(t|t_1) = H\left(\frac{N(t_1,t)}{Q(t)}\right)\frac{\psi}{Q(t)} = H(0)\frac{\psi}{Q(t)} < \frac{\psi}{Q(t)}$ $(t_1 \leq t)$.

5.2 The no residual queue property

Recall that t_1 is the time of the last arrival to 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 at time t_1, $Q(t_1) > 0$, so that $\tau_1 > t_1$.

In the case of no queue priority we noted in Proposition 2 that 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 specify 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* $\tau_1 = t_1 + Q(t_1)/\psi$.

This condition implies that 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). He will be better off under any loose queue priority than under strict queue priority, since under the latter priority, he will arrive at τ_1 with certainty.

Proposition 4 *Under loose queue priority, Condition 2 implies the no residual queue property in Nash equilibrium.*

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 2. 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. ■

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

6 Concluding remarks

This paper considers bottleneck congestion where the arrival rate to the bottleneck is endogenous and user preferences are 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 a deterministic FIFO queue to a pure random queue. The paper defines the no residual queue property, which holds when the queue has vanished at the time of the last arrival. Proposition 1 shows that, independently of the queueing regime, the no residual queue property implies that the interval of arrival as well as the expected utility of users are uniquely determined in Nash equilibrium. The equilibrium expected utility is decreasing and concave as a function of the number of users, such that there is a congestion externality which increases in the number of users. The remainder of the paper then establishes that the no residual queue property holds in Nash equilibrium under all queueing regimes considered. Hence the results of Proposition 1 are robust with respect to random sorting of the queue.

This paper is structured around Proposition 1, which leaves open the characterisation of Nash equilibrium in situations when the no residual queue property does not hold. While we have not obtained general results, the situation with no queue priority and linear utility when $\alpha < \gamma$ is straightforward to analyse. The residual queue property does not hold, but the analysis in section 4 can still be extended to find the equilibrium arrival rate. In this case, arrivals to the queue begin at time $-\frac{N}{\psi}\frac{\alpha+\gamma}{\alpha+\gamma+2\beta}$. From that time until time 0 users arrive at the constant rate $\psi\frac{\alpha+\beta}{\alpha}$. Not all users have arrived before time 0 and the remaining users all arrive at time 0, such that the size of the queue jumps discontinuously to $N\frac{2\beta}{\alpha+\gamma+2\beta}$ at this time.

A potential direction for future research is the exploration of the properties of Nash equilibrium in networks of bottlenecks with loose or no queue priority. Bottlenecks may be interpreted as congested nodes in a network or they may represent

competing sequences of activities. In the case of strict queue priority bottlenecks in series, Nash equilibrium is determined by the bottleneck with the smallest capacity and so the analysis is essentially unchanged relative to the case of a single bottleneck. Similarly, under the no residual queue property, bottlenecks in parallel may be treated as just one bottleneck. When queueing regimes are allowed to be random, the former reduction is no longer valid. The latter reduction is not valid when the no residual queue property does not hold. A network of bottlenecks may therefore have different equilibrium properties than single bottlenecks. The properties of equilibrium in such networks remain to be explored. A first setting that could be explored is the combination of two parallel bottlenecks, one with NQP and one with SQP, when users are risk averse and the no residual queue property does not hold.

Another extension to consider is bottleneck congestion in a monocentric city, where users are heterogenous in the distance to the city centre. It is feasible to use our general formulation of preferences in such an analysis. The formulation of the no residual queue property must be adapted to the fact that the first and the last user will not be identical. Under strict queue priority it turns out that users endogenously sort in time, such that the most distant users depart early and arrive late at the city centre. The properties of equilibrium under random queue sorting will be interesting to explore. Finally, we like to point to the range of open questions that arise when urban structure is allowed to be endogenous. Ultimately, we seek an equilibrium theory that incorporates congestion and risk as well as individual decisions concerning location in both time and space.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

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

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

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

[8] Gross, D., Shortle, J. F., Thomson, J. M. and Harris, C. (2008) *Fundamentals of Queueing Theory* fourth edn John Wiley & Sons Hoboken, NJ.

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

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

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

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

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

[14] Newell, G. F. (1982) *Applications of Queueing Theory* 2nd. edn Chapman & Hall.

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

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

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

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