The Concert Queueing Game: Strategic Arrivals with Waiting and Tardiness Costs

Sandeep Juneja* Nahum Shimkin†

August 21, 2012

Abstract

We consider the non-cooperative choice of arrival times by individual users, who seek service at a first-come first-served queueing system that opens up at a given time. Each user wishes to obtain service as early as possible, while minimizing the expected wait in the queue. This problem was recently studied within a simplified fluid-scale model. Here we address the unscaled stochastic system, assuming a finite (possibly random) number of homogeneous users, exponential service times, and linear cost functions. In this setting we establish that there exists a unique Nash equilibrium, which is symmetric across users, and characterize the equilibrium arrival-time distribution of each user in terms of a corresponding set of differential equations. We further establish convergence of the Nash equilibrium solution to that of the associated fluid model as the number of users is increased. We finally consider the price of anarchy in our system and show that it exceeds 2, but converges to this value for a large population size.

1 Introduction

The so-called concert queueing game, presented in [9], addresses the strategic choice of arrival times to a service facility that opens up at a given time, and serves its users (or

*School of Technology and Computer Science, Tata Institute of Fundamental Research, Mumbai, India. email: juneja@tifr.res.in

†Department of Electrical Engineering, Technion – Israel Institute of Technology, Haifa 32000, Israel. e-mail: shimkin@ee.technion.ac.il

customers) according to the order of their arrival to the queue. Users wish to conclude their service as early as possible, while minimizing their wait in the queue. Accordingly, the cost function of each user is composed of a penalty term for late completion times, or tardiness, and a penalty term for long waits in the queue, where both terms are taken to be linear in their variables. The motivation for this model comes from queues that form in front of concert and movie theater box offices, in banks before opening, in cafeterias before lunch, in shops upon the launching of a popular gadget, etc. The above-mentioned paper considers a simplified fluid model, where the users are points in a continuum. This model was extended in [10] to multiple classes of users, which differ in their cost parameters, and in [8] to networks of parallel queues. Our goal here is to extend the analysis of the basic fluid model to a finite-population stochastic queueing system. We shall restrict attention to a homogeneous (single-class) user population, and a single-server queue.

Fluid models that address the strategic choice of arrival time to a shared facility have been extensively treated in the transportation literature, starting with the seminal formulation of the bottleneck model in [20]. In this model, also known as the morning commute problem, it is usually assumed that travelers have a target time at which they wish to get to their destination, and are accordingly penalized for being too early or too late. For details see [16, 13] and references therein. We note that the (multiclass) fluid model in [10] can be seen as special cases of this class of models, where the users prefer to complete their service as *early* as possible. The main contribution being the explicit derivation the equilibrium solution and some of its properties for this model.

The strategic choice of arrival times into queues with a limited service period were apparently first considered in [3], where a Poisson-distributed number of homogeneous users with exponential service requirements arrive at a service facility with known opening and closing time, and wish to minimize their own waiting time. This work showed that the arrival profile in the symmetric equilibrium is a continuous probability distribution that extends over a finite interval before and after the opening time, and derived a set of differential equations that characterizes it. Several variations of this model have since been considered. The recent paper [6] analyzes the same model except that queueing before the opening time is not allowed. Consequently, the equilibrium arrival distribution includes a point mass at the opening time. The related work in [18, 19] studies models with discrete arrival instances and deterministic service durations. Their theoretical predictions were compared to empirical finding in controlled laboratory experiments, which provide support for the symmetric mixed-strategy equilibrium solution on the aggregate level (although not on the individual level). More details on these papers and other related work can be found in [6]; see also [5] (Chapter 6).

The above-mentioned queueing models did not incorporate an explicit tardiness penalty in the cost function, but rather assumed a known closing time which provides the incentive for users to arrive early (or at least not too late). The recent paper [7] studies several variants of the queueing and fluid model with the same cost function as in [10]. Similar to [3] and [6], a Poisson number of homogeneous arrivals is assumed, and the symmetric equilibrium is analyzed. Both versions with and without early arrivals are considered, along with their fluid analogues.

In the present paper we focus on the stochastic queueing model with tardiness costs, exponential service, early arrivals, and no closing time. We note that the latter two model assumptions are made here for concreteness and brevity, and the essential part of the analysis should carry over to other variants of the model that modify these assumptions. Our main contributions with respect to the above-mentioned literature are the following.

1. *General population size:* We consider a general distribution of N, the number of customers, rather than restrict attention to a Poisson-distributed number. To this end, we first address in detail the model with deterministic N, and then generalize on this basis to an arbitrary distribution. This extension requires, in particular, to modify the differential equations that describe the symmetric equilibrium.

2. *Equilibrium analysis:* We provide a more complete analysis of the existence and uniqueness of the equilibrium arrival profile. While previous work focused at the outset on symmetric arrival distributions (i.e., identical for all users), we first establish that any equilibrium must be symmetric (under a mild technical condition). We further provide a rigorous analysis existence and uniqueness of solutions to the differential equation that defines the symmetric equilibrium.

3. *Convergence to the fluid model solution:* We show that as N increases to infinity, the equilibrium arrival profile converges to the solution of the fluid model in [10], and provide bounds for the rate of convergence.

4. *Price of Anarchy (PoA)*: We show that the PoA (defined as the ratio of the social cost of the worst-case non-cooperative equilibrium to the optimal social cost) in our stochastic model is always larger than 2, and converges to this value as N increases to infinity.

It should be emphasized that non-existence of asymmetric equilibria, as established here, cannot be taken for granted. As is well known, multiple non-symmetric equilibria may arise even in the simplest symmetric games (e.g., consider two-person coordination game with payoff matrix whose rows are $[a, 0]$ and $[0, b]$, $a, b > 0$, identical for both players). More specifically, in the single-class fluid model studied in [10], while the aggregate arrival profile of the user population is uniquely determined, the individual user decisions are not. Indeed, the (continuous) user population may be split into any number of groups, with one arriving before the other.

The results on convergence of the equilibrium solution to that of the fluid model are apparently new for the considered class of arrival-time queueing games. As shown in [10], the fluid model is extremely useful in providing simple and elegant equilibrium solutions to a variety of what-if perturbations and associated optimizations to the concert queuing game. Our convergence results lend credibility to such analysis, at least when a large number of customers are involved. In the transportation literature, the relation between the extensively-studied fluid bottleneck model and a finite population model has been studied numerically, e.g., in [17]. However, we are not aware of analytical studies of convergence to the fluid model. Asymptotic (diffusion and fluid scale) analysis of strategic equilibrium in queues with rational customers has been the subject of several recent papers, including [2, 14, 11, 1], however the models there are quite different from ours and mainly pertain to discrete arrival and routing decisions. In another direction, our results may be related to the evolving discipline of Mean-Field Games [12], which studies the limit of N-player dynamic games as N increases to infinity. Our model may be considered as a middle ground between static (normal-from) games and fully dynamic games, in the sense that, while the temporal aspect is certainly present, each player takes a single decision prior to the start of the game.

The paper is organized as follows. Section 2 describes the basic model and defines the relevant notion of a Nash equilibrium profile. Section 3 presents some relations and preliminary results related to the queue process and user costs. In Section 4 we establish that any equilibrium profile must be user-symmetric, characterize the equilibrium arrival distribution in terms of a set of differential equations, and establish existence and uniqueness of the equilibrium profile. Section 5 addresses the computation of the equilibrium distribution. Explicit expressions are derived for the case of $N = 2$ (two-user system), while for the general case a numerical procedure is suggested, and the results illustrated for different values of N. In Section 6 we outline the extension to the random-N model. Section 7 establishes the convergence of the equilibrium profile to the corresponding fluid solution, while Section 8 addressed the PoA for this model. We conclude the paper in Section 9 with a brief summary and future research directions.

To facilitate the presentation, some of the more technical proofs in Sections 3 and 4 are relegated to the Appendices.

2 Model Description

We consider a service system that caters to arriving users on a first-come first-served basis. The user population is finite and of size $N \geq 2$, which is taken for the time being to be a deterministic number. Thus, each user supposes that there are $M = N - 1$ *other* users that are due to arrive[1]. Service starts at time $t = 0$, and proceeds until all users are served. The required service times of all users are independent and exponentially distributed. Customers may arrive and queue up both before and after the opening time. Thus, each user $i \in \{1, \ldots, N\}$ may choose his arrival time t_i, possibly randomly according to some probability distribution on the real line, with cumulative distribution function (CDF) $F_i(t)$. If several users arrive simultaneously, they are randomly ordered.

When choosing their arrival times, users weigh the benefit of early service completion with the cost of a long wait in the queue. Suppose user i arrives at time t_i, waits in the queue for w_i time units, and enters service at time τ_i (hence $w_i = \tau_i - t_i$). His cost function is then

$$c_i(w_i, \tau_i) = \alpha w_i + \beta \tau_i$$

where $\alpha > 0$, $\beta > 0$ are the respective cost sensitivities. Thus, we focus here on linear costs. We further suppose that users are homogeneous in terms of their cost functions, so that α and β are identical for all users.

An *arrival profile* $\mathcal{F} = \{F_i\}$ is a collection of arrival time distributions of all users. Given $\mathcal{F}$ and the above system description, both w_i and τ_i may be seen to be well-defined random variables, and we may consider the expected cost

$$C_i(\mathcal{F}) = E_{\mathcal{F}}(\alpha w_i + \beta \tau_i)\,,$$

where $E_{\mathcal{F}}$ is the expectation induced by $\mathcal{F}$.

As usual, we say that the arrival profile $\mathcal{F} = \{F_i\}$ is a Nash equilibrium if no user can improve his expected cost by a unilateral change of his arrival time distribution. Formally,

[1] The relation between N and M is more intricate in the stochastic case, as discussed in Section 6.

Definition 1 *An arrival profile* $\mathcal{F} = \{F_i\,,\ i = 1, \ldots, N\}$ *is a* Nash equilibrium point (NEP) *if*

$$C_i(\mathcal{F}) \leq C_i(\tilde{F}_i, \mathcal{F}^{-i}) \tag{1}$$

for every user i *and every CDF* $\tilde{F}_i$ *on the real line. Here* $(\tilde{F}_i, \mathcal{F}^{-i})$ *stands for the profile* $\mathcal{F}$ *with* F_i *replaced by* $\tilde{F}_i$*.*

Our main objective is to characterize the NEPs and study their properties.

A somewhat more explicit characterization of the equilibrium will be useful. Let $C_i(t, \mathcal{F}^{-i})$ denote the expected cost of user i if he joins the queue at time t, while all others follow their arrival distributions as specified in $\mathcal{F}^{-i}$. Evidently

$$C_i(\mathcal{F}) \equiv C_i(F_i, \mathcal{F}^{-i}) = \int C_i(t, \mathcal{F}^{-i}) dF_i(t)\,.$$

It is now easily seen that the NEP can be characterized as follows.

Lemma 1 *An arrival profile* $\mathcal{F} = \{F_i\}$ *is a Nash equilibrium point if, and only if, for every user* i *there exists a constant* c_i *so that*

(i) $C_i(t, \mathcal{F}^{-i}) \geq c_i$ *for all* t*.*

(ii) $C_i(t, \mathcal{F}^{-i}) = c_i$ *on a set* A_i *of* F_i*-measure 1, namely* $\int_{A_i} dF_i(t) = 1$*.*

We refer to c_i as the *equilibrium cost* for user i, at a given equilibrium point.

Recall that the *support* of a probability measure is defined as the smallest closed set of probability 1. Let $\mathcal{T}_i$ denote the support of F_i (i.e., of the measure represented by F_i). The following technical assumption will henceforth be imposed on the arrival time distributions.

Assumption 1 *For each* F_i*, the corresponding support* $\mathcal{T}_i$ *can locally (i.e., on any finite interval) be represented as a finite union of closed intervals and points.*

This assumption rules out elaborate distributions that are supported on an infinite number of distinct components over a finite span. Such elaborate constructs can arguably be ruled out as reasonable arrival distributions even for the mathematically-inclined user. The assumption will be used in Lemma 15 and Proposition 1 to establish uniqueness of the equilibrium profile. We conjecture that uniqueness still holds without this assumption, but currently have no proof for that.

3 Preliminaries

3.1 Basic Queueing Relations

We briefly recall some relevant queueing relations for our system, and establish some useful relations between the arrival profile and the queue size. Fix an arrival profile $\mathcal{F} = \{F_i\}$. The cumulative arrival process $\mathbf{A}(t)$ can be expressed as $\mathbf{A}(t) = \sum_{i=1}^{N} 1_{\{T_i \le t\}}$, where the T_i's are independent random variables, distributed as $T_i \sim F_i$. Clearly,

$$E(\mathbf{A}(t)) = \sum_i F_i(t) \overset{\triangle}{=} F(t)\,,$$

where $F(t) = \sum_i F_i(t)$ is the *aggregate* arrival distribution.

Let $(V_i : i \ge 1)$ denote the i.i.d., exponentially distributed sequence with mean μ^{-1}, where V_i for $i \le N$ is the service time of the user which is the i-th to be served. Let $\mathbf{Q}(t)$ denote the number of users in the queue at time t (including the one in service). For any set C, let $1_{\{C\}}$ denote its indicator function. Further define the following processes:

$$\begin{aligned}
&\mathbf{S}(t) = 1_{\{t\ge0\}} \max_{m\ge0}\{\sum_{i=1}^{m} V_i \le t\}\\
&\mathbf{B}(t) = 1_{\{t\ge0\}} \int_0^t 1_{\{\mathbf{Q}(s)>0\}} ds\\
&\mathbf{I}(t) = t1_{\{t\ge0\}} - \mathbf{B}(t) = 1_{\{t\ge0\}} \int_0^t 1_{\{\mathbf{Q}(s)=0\}} ds\,.
\end{aligned}$$

Here $\mathbf{S}(t)$ denotes the number of service completions if the server is busy for $t \ge 0$ time units (recalling that service commences at $t = 0$) and there are infinitely many users (we are not restricting $\mathbf{S}(t)$ to N simply for notational convenience), $\mathbf{B}(t)$ denotes the time that the queue has been busy up to time t, and $\mathbf{I}(t)$ is the idle time process. The queue length process satisfies the following sample path equality: $\mathbf{Q}(t) = \mathbf{A}(t) - \mathbf{S}(\mathbf{B}(t))$. According to these definitions, all processes are continuous on the right with left limits.

Taking expected values, we obtain

$$\begin{aligned}
E(\mathbf{S}(t)) &= \mu t 1_{\{t\ge0\}} && (2)\\
E(\mathbf{I}(t)) &= 1_{\{t\ge0\}} \int_0^t P\{\mathbf{Q}(s) = 0\} ds && (3)\\
Q(t) \overset{\triangle}{=} E(\mathbf{Q}(t)) &= F(t) - \mu E(\mathbf{B}(t)) && (4)\\
&= F(t) - \mu t 1_{\{t\ge0\}} + \mu E(\mathbf{I}(t))\,.
\end{aligned}$$

We next point out a few properties of the queue process that will prove useful in the sequel. The following lemma relates the discontinuities of the expected queue size $Q(t)$ to those of the aggregate arrival profile $F(t)$.

Lemma 2 *Let $F(t-)$ denote the left limit of F at t, and similarly for $Q(t-)$. For every t,*

$$Q(t) - Q(t-) = F(t) - F(t-)\,.$$

That is, $Q(t)$ is continuous where $F(t)$ is, and at points of discontinuity of F, Q has upward jumps of equal magnitude.

Proof: The claim follows immediately from (4), once we note that $\mathbf{B}(t)$ is a continuous process. □

The next result addresses the monotonicity of the queue size in the arrival distribution.

Lemma 3

(i) Let $\mathcal{F} = \{F_i\}$ and $\tilde{\mathcal{F}} = \{\tilde{F}_i\}$ be two arrival profiles, and let $Q(t)$ and $\tilde{Q}(t)$ be the respective expected queue sizes. Suppose $\mathcal{F}$ dominates $\tilde{\mathcal{F}}$ up to some time t, in the sense that $F_i(s_2) - F_i(s_1) \geq \tilde{F}_i(s_2) - \tilde{F}_i(s_1)$ for all $s_1 < s_2 \leq t$ and all i. Then $Q(t) \geq \tilde{Q}(t)$.

(ii) Furthermore, if strict inequality holds for some $s_1 < s_2 \leq t$ and i, then $Q(t) > \tilde{Q}(t)$.

(iii) Assertions (i) and (ii) continue to hold when $Q(t)$ is replaced by the probability $P\{\mathbf{Q}(t) > 0\}$, and $\tilde{Q}(t)$ by $P\{\tilde{\mathbf{Q}}(t) > 0\}$.

Proof: The claim follows by a stochastic coupling argument, see Appendix A. □

The following continuity result is key to the existence and uniqueness of solutions to the differential equation (12), which will describe the equilibrium distribution. Let

$$\|f\|_t = \sup_{0 \leq s \leq t} |f(s)|$$

denote the sup-norm of a real function f on $[0,t]$.

Lemma 4 *$P\{\mathbf{Q}(t) = 0\}$ is Lipschitz continuous in the arrival distribution, in the sense that there exists a constant $K > 0$ such that*

$$|P\{\mathbf{Q}(t) = 0\} - P\{\tilde{\mathbf{Q}}(t) = 0\}| \leq K \sum_{i=1}^{N} \|F_i - \tilde{F}_i\|_t$$

for any pair of arrival profiles $\mathcal{F} = (F_i)$ and $\tilde{\mathcal{F}} = (\tilde{F}_i)$, and all $t \geq 0$.

Proof: The proof of this lemma again relies on stochastic couplings between the perturbed systems. It is given in Appendix A. □

We next consider the evolution of the queue length probabilities $P\{\mathbf{Q}(t) = k\}$, under a given arrival profile $\mathcal{F}$. Some care is required in writing the evolution equations since the arrival distributions are not memoryless, and the arrival intensity depends on which users already arrived. Let $\mathcal{N}(t) \in 2^{\{1,\dots,N\}}$ denote the set of users that arrived up to time t (inclusive). Let $p_t(k, \mathcal{N}) = P\{\mathbf{Q}(t) = k, \mathcal{N}(t) = \mathcal{N}\}$ denote the probability that at time t there are k users in the queue and the set $\mathcal{N}$ of users have arrived. Suppose for the time being that F_i admits a density F_i', and recall that $H_i(t) = F_i'(t)/(1 - F_i(t))$ is hazard rate function associated with F_i. It is now easily verified that the pair $(\mathbf{Q}(t), \mathcal{N}(t))$ forms a Markov chain, with flow balance equations as summarized in the following lemma.

Lemma 5 *Suppose $F_i(t_0) = 0$ for some t_0 small enough and for all i. Then*

$$\begin{aligned}\frac{d}{dt}p_t(k,\mathcal{N}) = &-\left(\mu 1_{\{t\geq 0,\, k>0\}} + \sum_{i\notin\mathcal{N}} \frac{F_i'(t)}{1-F_i(t)}\right) p_t(k,\mathcal{N}) \\ &+ 1_{\{k>0\}} \sum_{i\in\mathcal{N}} \frac{F_i'(t)}{1-F_i(t)} p_t(k-1, \mathcal{N}\setminus\{i\}) \\ &+ \mu 1_{\{t\geq 0,\, k<N\}} p_t(k+1,\mathcal{N})\end{aligned} \tag{5}$$

for all $0 \leq k \leq N$, $\mathcal{N} \in 2^{\{1,\dots,N\}}$ and $t \geq t_0$, with initial conditions $p_{t_0}(0, \emptyset) = 1$.

Remark 1 *The equations above can be easily modified when the density of $F_i(t)$ is not well defined everywhere, by writing $dF_i(t)$ in place of $F_i'(t)dt$.*

The following lemma establishes that the probability of an empty queue is increasing when there are no arrivals.

Lemma 6 *Suppose $\frac{d}{dt}F(t) = 0$ (equivalently $\frac{d}{dt}F_i(t) = 0$ for all i) for some $t \geq 0$, where $\frac{d}{dt}$ denotes here the right-hand derivative. Then $\frac{d}{dt}P\{\mathbf{Q}(t) = 0\} \geq 0$, with strict inequality if $P\{\mathbf{Q}(t) = 0\} < 1$ (equivalently, if $F(t) > 0$).*

Proof: Since service in on from $t = 0$ and $F_i'(t)$ is the arrival density at time t, then with $F_i'(t) = 0$ for all i we obtain from Lemma 5,

$$\frac{d}{dt}P\{\mathbf{Q}(t) = 0\} = \mu P\{\mathbf{Q}(t) = 1\},$$

which is strictly positive if $P\{\mathbf{Q}(t) = 1\} > 0$. Now, $P\{\mathbf{Q}(t) = 0\} < 1$ implies (and is implied by) positive probability for arrivals up to t, namely $F(t) > 0$, in which case there is positive probability for one user remaining in the queue, or $P\{\mathbf{Q}(t) = 1\} > 0$. □

3.2 Cost Properties

Consider a user who arrives at time t and sees a queue of size q (including the one in service) ahead of him. To enter service, he needs to wait till these q users get served, and if $t < 0$ he needs to wait in addition $(-t)$ time units till service starts. Therefore, his expected waiting time is given by $E(w) = \mu^{-1}q - t1_{\{t<0\}}$. As this user will enter service at $\tau = w + t$, we also have $E(\tau) = \mu^{-1}q + t1_{\{t\geq 0\}}$. Therefore, the expected cost of this user would be

$$\begin{aligned} E(\alpha w + \beta\tau) &= \alpha(\mu^{-1}q - t1_{\{t<0\}}) + \beta(\mu^{-1}q + t1_{\{t\geq 0\}}) \\ &= (\alpha+\beta)\mu^{-1}q - \alpha t1_{\{t<0\}} + \beta t1_{\{t\geq 0\}}\,. \end{aligned}$$

Given an arrival profile $\mathcal{F}$, recall that $Q(t)$ denotes the expected queue size at time t, and let $Q^{-i}(t)$ denote the expected queue size *with the arrival of user i excluded.* Evidently $Q^{-i}(t)$ depends on $\mathcal{F}^{-i}$ only; it will play a central role in determining user i's cost. Consider a potential arrival of user i at time t. If $Q^{-i}(t)$ is continuous at t, then the expected number of users that i will see before him is precisely $Q^{-i}(t)$, and his expected cost for arriving at t would be

$$C_i(t, \mathcal{F}^{-i}) = \frac{\alpha+\beta}{\mu} Q^{-i}(t) - \alpha t1_{\{t<0\}} + \beta t1_{\{t\geq 0\}}\,. \tag{6}$$

If $Q^{-i}(t)$ has an upward jump at t (due to a corresponding jump in F^{-i}, see Lemma 2), then the expected number of users before i would be $\bar{Q}^{-i}(t) = \frac{1}{2}(Q^{-i}(t-) + Q^{-i}(t))$. The above expression still holds with Q^{-i} replaced by $\bar{Q}^{-i}$. Thus, in general,

$$C_i(t, \mathcal{F}^{-i}) = \frac{\alpha+\beta}{\mu} \frac{Q^{-i}(t-) + Q^{-i}(t)}{2} - \alpha t1_{\{t<0\}} + \beta t1_{\{t\geq 0\}}\,. \tag{7}$$

The following observations are immediate from the last expression and Lemma 2.

Lemma 7 *$C_i(t, \mathcal{F}^{-i})$ is continuous in t, except at points where $Q^{-i}(t)$ has an (upward) jump. These discontinuity points are, equivalently, the points where $F^{-i} = \sum_{j\neq i} F_j$ has a point mass. At points of discontinuity, $C_i(t, \mathcal{F}^{-i})$ is continuous from the right and has left limit, and $C_i(t-, \mathcal{F}^{-i}) < C_i(t, \mathcal{F}^{-i})$ (i.e., upward jumps only).*

4 Existence, Uniqueness and Structure

We establish in this section the existence and uniqueness of the equilibrium profile. We first observe that the equilibrium profile must be symmetric, namely identical for all users. We will denote by G the arrival distribution in a symmetric equilibrium, so that $F_i \equiv G$ and $\mathcal{F} = G^N$. Our main results identify general structural properties of G, and fully characterize it as the solution of a differential equation with boundary value conditions.

The key finding of this section are summarized in the following theorem. Recall that $M = N - 1$, the number of other users as seen by any single user.

Theorem 1

(i) The equilibrium profile $\mathcal{F}$ exists, is unique, and has the symmetric form $\mathcal{F} = G^N$.

(ii) The support $\mathcal{T}_G$ of G is a single finite interval, denoted $[t_a, t_b]$, with $t_a < 0$ and $t_b > 0$.

(iii) $G(t)$ is continuous, and admits a right-derivative $G'(t)$ for all t. Further, $G'(t)$ is continuous (hence coincides with the full derivative) everywhere except at $t = t_a$ and $t = 0$.

(iv) For $t_a \le t < 0$, $G'(t)$ is constant and given by $\frac{M}{\mu} G'(t) = \frac{\alpha}{\alpha+\beta}$.

(v) For $0 \le t \le t_b$, $G'(t)$ satisfies the functional differential equation (FDE)

$$\frac{M}{\mu} G'(t) = \frac{\alpha}{\alpha + \beta} - P\{\mathbf{Q}^{-i}(t) = 0\}, \tag{8}$$

where $\mathbf{Q}^{-i}$ is the queue length process that corresponds to $\mathcal{F}^{-i} = G^M$.

(vi) $G(t)$ obeys the terminal conditions $G(t_b) = 1$ and $G'(t_b) = 0$.

(vii) Properties (ii)-(vi) above uniquely determine the equilibrium distribution G

Figure 1 illustrates the structure of G as outlined above. We note that G' has a downward jump of magnitude $\frac{\mu}{M} P\{\mathbf{Q}^{-i}(0) = 0\} > 0$, as follows from (iv) and (v).

The rest of this section is devoted to the proof of the last theorem. We start the analysis in the next subsection by showing that any equilibrium profile must be symmetric. Focusing on the symmetric case thereafter, in Subsection 4.2 we show that G^N is an equilibrium profile if and only if the distribution function G satisfies properties (ii)-(vi) of Theorem 1. Finally, in Subsection 4.3 we show that such a G exists and is unique.

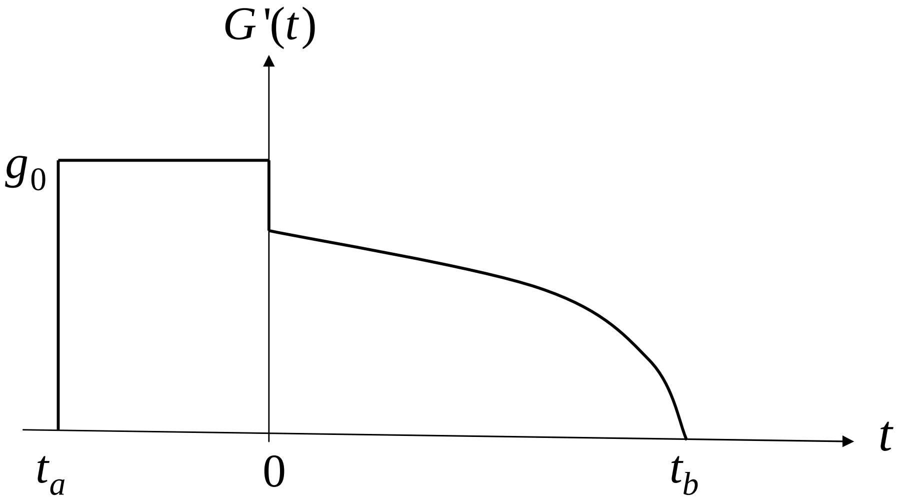

Figure 1: Sketch of the equilibrium arrival-time density.

4.1 Symmetry

We first establish that any equilibrium profile must be symmetric.

Proposition 1 *Any equilibrium profile $\mathcal{F} = \{F_i\}$ is symmetric, in the sense that $F_i = F_j$ for all i, j. Furthermore, the F_i's do not contain any point masses.*

The proof proceeds through several lemmas, and is presented in Appendix B. Essentially, after establishing some auxiliary claims in Lemmas 12 and 13, we show in Lemma 14 that the equilibrium cost c_i of all users must be identical, as otherwise users with inferior costs could improve their positions. Observing the cost expressions in equations 6-7, we infer from Lemma 15 that the *expected* queue sizes $Q^{-i}(t)$ as view by all users must be identical, which leads to identity of the arrival profiles. We mention that the last step, in particular, is far from trivial, and our proof of the latter lemma is intricate and nuanced.

4.2 Characterization

Let us fix some notation. Given an arrival profile $\mathcal{F} = \{F_j\}_{j=1}^N$, recall that $\mathcal{F}^{-i}$ denotes the same collection with F_i excluded, and $F^{-i} = \sum_{j \neq i} F_j$. Let $\mathbf{Q}^{-i}(t)$ denote that queue-length process that corresponds to $\mathcal{F}^{-i}$. Evidently, for a symmetric arrival profile with $F_i \equiv G$ we obtain $F^{-i} = MG$, and $\mathbf{Q}^{-i}(t)$ does not depend on i. However, it will be convenient to retain this notation with a generic index i. We shall write $C_i(t)$ for $C_i(t, \mathcal{F}^{-i})$ with

$\mathcal{F}^{-i} = G^M$. The common value of equilibrium costs c_i is denoted by c_0.

Lemma 8 *Let $\mathcal{T}_G^o$ denote the interior of the support of G. Then*

$$\frac{M}{\mu} G'(t) = \frac{\alpha}{\alpha + \beta} - P\{\mathbf{Q}^{-i}(t) = 0\} 1_{\{t \geq 0\}}\,, \quad t \in \mathcal{T}_G^o\,. \tag{9}$$

Here G' denotes the derivative of G, which exists and is continuous for all $0 \neq t \in \mathcal{T}_G^o$. For $t = 0$, $G'(0)$ refers to the right-derivative.

Proof: As G has no point masses (Proposition 1), then $Q^{-i}(t)$ is continuous by Lemma 7, so that Equation (6) is in effect. Using the relations in (4) and (3), along with $F^{-i} = MG$, obtains

$$\begin{aligned} C_i(t) &= (\alpha + \beta) \left[\mu^{-1} F^{-i}(t) + E(\mathbf{I}^{-i}(t))\right] - \alpha t \\ &= (\alpha + \beta) \left[\frac{M}{\mu} G(t) + 1_{\{t \geq 0\}} \int_0^t P\{\mathbf{Q}^{-i}(s) = 0\} ds\right] - \alpha t\,. \end{aligned}$$

Therefore,

$$\frac{d}{dt} C_i(t) = (\alpha + \beta) \left[\frac{M}{\mu} G'(t) + P\{\mathbf{Q}^{-i}(t) = 0\} 1_{\{t \geq 0\}}\right] - \alpha \tag{10}$$

wherever the derivatives exist (which is almost everywhere since G is monotone). Now, by Lemma 13, $C_i(t)$ is constant on $\mathcal{T}_i = \mathcal{T}_G$, so that $\frac{d}{dt} C_i(t) = 0$ in $\mathcal{T}_G^0$, and (9) follows. Further, $P\{\mathbf{Q}^{-i}(t) = 0\}$ is continuous in t (again, since G contains no point masses), so that the right-hand side of (9) is continuous for all t, except for a possible discontinuity at $t = 0$ due to the indicator function. This establishes the claim regarding pointwise existence and continuity of $G'(t)$. □

Proposition 2 *Let $\mathcal{F} = G^N$ be an equilibrium arrival profile. Then G satisfies properties (ii)-(vi) of Theorem 1.*

Proof: The enumeration below refers to properties (ii)-(vi) of Theorem 1.

(ii) Boundedness of the support was established in Lemma 13. Recall that the support is closed by definition. To show that it consists of a single interval, we thus need only show that it has no "gaps". Consider $t_1 < t_2 < t_3$ so that $[t_1, t_2]$ is contained in $\mathcal{T}_G$, while (t_2, t_3) is not. We will show that t_3 cannot be in the support. Since (t_2, t_3) is not in $\mathcal{T}_G$ we have $G'(t) = 0$ there, so that (10) reduces to

$$\frac{d}{dt} C_i(t) = (\alpha + \beta) P\{\mathbf{Q}^{-i}(t) = 0\} 1_{\{t \geq 0\}} - \alpha\,. \tag{11}$$

Further, by Lemma 1, $C_i(t) \equiv c_0$ on $[t_1, t_2]$, while $C_i(t) \geq c_0$ on (t_2, t_3). Therefore $\frac{d}{dt}C_i(t_2+) \geq 0$, implying that $P\{\mathbf{Q}^{-i}(t_2) = 0\}1_{\{t_2 \geq 0\}} - \alpha \geq 0$. Clearly this cannot hold for $t_2 < 0$ since $\alpha > 0$. Suppose then that $t_2 \geq 0$. As there are no arrivals on (t_2, t_3), then $P\{\mathbf{Q}^{-i}(t) = 0\}$ is strictly increasing there (due to departures; notice that $P\{\mathbf{Q}^{-i}(t_2) = 0\} < 1$ due to arrivals on $[t_1, t_2]$), so that $\frac{d}{dt}C_i(t) > 0$ there. Hence $C_i(t_3) > C_i(t_2) = c_0$, implying that t_3 is not in the support of the equilibrium distribution G. It follows that $\mathcal{T}_G$ indeed consists of a single interval.

It remains to show that $t_a < 0$ and $t_b > 0$. Observe that c_0, the common equilibrium cost, is strictly larger than 0 (as no wait for all is impossible). Suppose $t_a \geq 0$. Then an arrival at $t = 0$ would incur zero cost, namely $C_i(0) = 0 < c_0$. But this contradicts the minimal cost property of the equilibrium (Lemma 1). Hence $t_a < 0$. Suppose next that $t_b < 0$. Since $(t_b, 0) \not\subset \mathcal{T}_G$ then (11) is in effect, and we obtain that $\frac{d}{dt}C_i(t) < 0$ there. This means that $C_i(t) < C_i(t_2) = c_0$ on $(t_b, 0)$, again contradicting the minimal cost property of equilibrium. Finally, suppose that $t_b = 0$. This means that (all) arrivals occur before $t = 0$ with probability 1, and as there is no service by that time it follows that $P\{\mathbf{Q}^{-i}(0) = 0\} = 0$. Then (11) again implies that $\frac{d}{dt}C_i(t) < 0$ for $t = 0+$, which conflicts with the minimal cost property as before.

(iii) Outside of $[t_a, t_b]$ we obviously have $G'(t) = 0$. Continuity of $G'(t)$ on $(t_a, t_b) \setminus \{0\}$ follows by Lemma 8. Continuity of $G'(t)$ at t_b is claimed in (vi), which is establishes subsequently.

(iv) and (v): Follow directly from (9), as $\mathcal{T}_G^o = (t_a, t_b)$.

(vi) That $G(t_b) = 1$ is obvious by definition of t_b. Consider $G'(t_b) = 0$. Let $G'(t_b-)$ denote the left limit of of $G'(t)$ at t_b, which exists as follows from (11). Suppose $G'(t_b-) > 0$. Observe that $G'(t_b+) = 0$ (as $t > t_b$ lies outside the support of G). Thus, by (10), $C_i'(t_b+) < C_i'(t_b-) = 0$ (the latter holds since $C_i(t) = c_0$ on (t_a, t_b)). But this implies that $C_i(t_b + \epsilon) < C_i(t_b) = c_0$, hence G cannot be an equilibrium distribution. □

Proposition 3 now establishes the converse to Proposition 2.

Proposition 3 *Suppose a probability distribution function $G(t)$ satisfies properties (ii)-(vi) of Theorem 1. Then $\mathcal{F} = G^N$ is an equilibrium profile.*

Proof: We will show that $C_i(t) = C_i(t, G^N)$ satisfies the requirements of Lemma 1. That $C_i(t) \equiv c$ (a constant) for $t \in [t_a, t_b]$ follows by (iv) and (v) by construction, as seen from (10). It remains to show that $C_i(t) \geq c$ for $t \notin [t_a, t_b]$. For $t < t_a < 0$, (10) implies that

$C_i'(t) = -\alpha < 0$, hence indeed $C_i(t) > C_i(t_b) = c$. Consider $t > t_b > 0$. At $t = t_b$ we have by (v) and (vi) that $P\{\mathbf{Q}^{-i}(t) = 0\} = \frac{\alpha}{\alpha+\beta}$. Further, since there are no arrivals for $t > t_b$ (as $G(t_b) = 1$), then $P\{\mathbf{Q}^{-i}(t) = 0\}$ is strictly increasing there. It then follows from (10) that for $t > t_b$, $C_i'(t) = (\alpha + \beta)P\{\mathbf{Q}^{-i}(t) = 0\} - \alpha > C_i'(t_b) = 0$. □

The last two propositions establish claims (ii)-(vii) of Theorem 1. These properties provide the general form of the symmetric equilibrium distribution $G(t)$, which we repeat here in a more explicit form:

1. $G(t) = 0$ up to some point $t_a < 0$.

2. For $t \in [t_a, 0)$, $G'(t) = g_0$ where $g_0 = \frac{\alpha}{\alpha+\beta}\frac{\mu}{M}$, a uniform distribution. Hence $G(t) = g_0(t - t_a)$. Note that t_a uniquely determines the equilibrium cost (and vice versa), via $c_0 = \alpha|t_a|$.

3. At $t = 0$, $G'(t)$ has a downward jump of magnitude $\frac{\mu}{M}P\{\mathbf{Q}^{-i}(0) = 0\}$. Since there is no service up to $t = 0$, we have $P\{\mathbf{Q}^{-i}(0) = 0\} = (1 - G(0))^M$.

4. For $t > 0$, $G(t)$ evolves according to the FDE (8), up to the point t_b where $G(t_b) = 1$ and $G'(t_b) = 0$.

It remains to establish existence and uniqueness of the equilibrium. Essentially, we will utilize monotonicity properties of the FDE (8), and show that there exists a unique t_a (equivalently, c_0) so that the solution $G(t)$ of that equation satisfies the required boundary conditions.

4.3 Existence and Uniqueness

We next show that there exists a unique probability distribution G that satisfies properties (ii)-(vi) of Theorem 1. In view of Propositions 2 and 3, this would imply the existence and uniqueness properties of the equilibrium as stated in claims (i) and (vii) of Theorem 1, and thereby complete the proof of that theorem.

We thus proceed to construct $G(t)$ that satisfied properties (ii)-(vi), and establish its uniqueness. For $t \in [t_a, 0)$, property (iv) implies that $G'(t) = g_0$ with $g_0 = \frac{\alpha}{\alpha+\beta}\frac{\mu}{M}$, hence $G(0) = |t_a|g_0$. We will henceforth consider $G(0) \in (0, 1)$ as a parameter. We then have $|t_a| = g_0/G(0)$, which determines $G(t)$ for $t < 0$.

Consider the differential equation (8) for $t \geq 0$. In general, $P\{\mathbf{Q}^{-i}(t) = 0\}$ is a function of $(G(s) : s \leq t)$. But since there is no service up to $t = 0$, the distribution of $\mathbf{Q}^{-i}(0)$ is fully determined by $G(0)$, as $P\{\mathbf{Q}^{-i}(0) = k\} = \binom{M}{k} G(0)^k (1 - G(0))^{M-k}$ for $0 \leq k \leq M$. In particular, $P\{\mathbf{Q}^{-i}(0) = 0\} = (1 - G(0))^M$. Therefore, we may consider $P\{\mathbf{Q}^{-i}(t) = 0\}$ as a function of $\mathbf{G_t} \overset{\triangle}{=} (G(s) : 0 \leq s \leq t)$ only. Let $P_0(\mathbf{G_t}) = P\{\mathbf{Q}^{-i}(t) = 0\}$ denote the probability of an arrival finding the queue empty at time t given $\mathbf{G_t}$. We thus obtain the FDE

$$\frac{M}{\mu} G'(t) = \frac{\alpha}{\alpha + \beta} - P_0(\mathbf{G_t}) \,, \tag{12}$$

with initial conditions $G(0)$. We aim to find a solution to this equation over an interval $[0, t_b]$ such that $G(t_b) = 1$ and $G'(t_b) = 0$ (conforming to property (vi) of Theorem 1), and show that such a solution is unique.

Let $\gamma_{\min} \in (0, 1)$ be the unique solution of $(1 - \gamma_{\min})^M = \frac{\alpha}{\alpha+\beta}$. Recalling that $P_0(\mathbf{G_0}) = (1 - G(0))^M$, it follows by (12) that $G'(0) < 0$ if $G(0) < \gamma_{\min}$. We therefore need only consider $G(0) \in [\gamma_{\min}, 1]$.

We next determine the interval $[0, \tau]$ on which (12) is defined. Recall that equation (12) is valid up to the point where $G(t) = 1$ (as there are no arrivals beyond that point). Further, the probability $P_0(\mathbf{G_t})$ is well defined as long as $G'(t) \geq 0$. Now, according to Lemma 6, if $G'(t_0) = 0$ at some point t_0 then $\frac{d}{dt} P_0(\mathbf{G_{t_0}}) > 0$ at that point, so that by (12) we obtain $G'(t) < 0$ for t beyond t_0, and the equation cannot be continued beyond that point. Thus, we need consider equation (12) for $0 \leq t \leq \tau$, where

$$\tau = \inf\{t \geq 0 : G(t) = 1 \text{ or } G'(t) = 0\} \,. \tag{13}$$

We shall refer to τ as the *final time* for the differential equation. We sometimes write $\tau(\gamma)$ for τ to make explicit the dependence on the initial conditions $G(0) = \gamma$.

Recall that

$$\|F - G\|_t = \sup_{0 \leq s \leq t} |F(s) - G(s)|,$$

It follows from Lemma 4 that $P_0(\mathbf{G_t})$ is Lipschitz continuous, in the sense that

$$|P_0(\mathbf{F_t}) - P_0(\mathbf{G_t})| \leq K \|F - G\|_t \tag{14}$$

for some $K > 0$ and all $t \geq 0$.

As a direct consequence we obtain some basic properties of the FDE.

Lemma 9

(i) The FDE (12) *with initial conditions* $G(0) \in [\gamma_{\min}, 1]$*, admits a unique solution* $G(t)$*,* $t \in [0, \tau]$*. Further, both* $G(t)$ *and* $G'(t)$ *are continuous in* t*.*

(ii) For fixed t*,* $G(t)$ *and* $G'(t)$ *are continuous functions of the initial condition* $G(0)$*. Further, continuity is uniform over finite time intervals.*

(iii) For fixed t*,* $G(t)$ *and* $G'(t)$ *are strictly increasing functions of the initial condition* $G(0)$*.*

Proof: See Appendix C. □

We next establish some key properties of the the final time τ. In particular, the following proposition shows that a transition from $G'(\tau) = 0$ to $G(\tau) = 1$ occurs at a unique value of $G(0)$.

Proposition 4 *There exists a unique* $\gamma^* \in (\gamma_{\min}, 1)$ *so that the solution* $G(t)$ *of the FDE* (12) *with initial condition* $G(0)$ *satisfies the following:*

(i) If $G(0) \in (\gamma^*, \gamma_{\min})$*, then* $G'(\tau) = 0$*,* $G(\tau) < 1$*, and* τ *is a strictly increasing function of* $G(0)$*.*

(ii) If $G(0) \in (\gamma^*, 1]$*, then* $G(\tau) = 1$*,* $G'(\tau) > 0$*, and* τ *is a strictly decreasing function of* $G(0)$*.*

Proof: See Appendix C. □

Proposition 4 shows that there can be at most one value $G(0) = \gamma^*$ for which both terminal conditions can hold simultaneously. A continuity argument is required to show that such a value does indeed exist. (We note that continuity of τ in $G(0)$ would suffice for that purpose. However, as τ is determined by level-crossing, its continuity is not straightforward from continuity of G and G', and a more refined argument is required.) This is taken up in the next theorem, which is main result of this section. Its proof is given in Appendix C.

Theorem 2 *There exists a unique number* $\gamma^* \in (\gamma_{\min}, 1)$ *so that the solution* $G(t)$ *of the differential equation* (12) *with initial condition* $G(0) = \gamma^*$ *satisfies both terminal conditions* $G(\tau) = 1$ *and* $G(\tau) = 0$*.*

5 Computation

In this section we demonstrate the explicit computation of the symmetric equilibrium distribution $G(t)$, or equivalently its density $g(t) = G'(t)$. For the two-user case ($N = 2$) we will be able to obtain closed-form solutions, while for $N > 2$ we will resort to numerical computation.

5.1 $N = 2$

Consider the case of two arriving users. Let $[t_a, t_b]$ denote the support of g, and recall that $t_a < 0$ and $t_b > 0$ by Theorem 1. For $t < 0$, by item (v) of that Theorem with $N = 2$ we obtain

$$g(t) = \mu\frac{\alpha}{\alpha+\beta}\,, \quad t \in [t_a, 0)\,. \tag{15}$$

For $t \geq 0$, by (8) we have

$$g(t) = \mu\frac{\alpha}{\alpha+\beta} - \mu P_0(t)\,,$$

where $P_0(t)$ denotes $P\{\mathbf{Q}^{-i}(t) = 0\}$ for notational ease. We will obtain an expression for $P_0(t)$ directly from the cost relations at equilibrium. Let c_0 be the equilibrium cost. Since the cost of the first arrival at t_a is $-\alpha t_a$, we have $c_0 = -\alpha t_a$. Now, observing the cost expression expression (6) and recalling that $C_i(t) = c_0$ on the support of g in equilibrium, we obtain

$$C_i(t) = \frac{\alpha+\beta}{\mu}Q^{-i}(t) + \beta t = c_0 = -\alpha t_a\,, \quad t \in [0, t_b]\,. \tag{16}$$

But for $N = 2$ each user i sees only one additional user in the system, so that $Q^{-i}(t) = P\{\mathbf{Q}^{-i}(t) = 1\} = 1 - P\{\mathbf{Q}^{-i}(t) = 0\}$. Therefore

$$P\{\mathbf{Q}^{-i}(t) = 0\} = 1 + \frac{\mu}{\alpha+\beta}(\beta t + \alpha t_a)\,,$$

a linear function of t. Thus,

$$\begin{aligned} g(t) &= \mu\frac{\alpha}{\alpha+\beta} - \mu P_0(t) \\ &= -\mu\frac{\beta}{\alpha+\beta} - \frac{\mu^2}{\alpha+\beta}(\beta t + \alpha t_a)\,, \quad t \in [0, t_b]\,. \end{aligned} \tag{17}$$

It remains to determine t_a and t_b. These can be obtained from (15) and (17) by using the terminal conditions $g(t_b) = 0$ and $G(t_b) = \int_{t_a}^{t_b} g(t)dt = 1$. After some computation, we obtain

$$-t_a = \mu^{-1}\sqrt{\frac{(2\alpha+\beta)\beta}{\alpha^2}} = \mu^{-1}\sqrt{\frac{\beta}{\alpha}(2+\frac{\beta}{\alpha})}$$
$$t_b = \frac{\alpha}{\beta}(-t_a) - \mu^{-1} = \mu^{-1}(\sqrt{1+\frac{2\alpha}{\beta}} - 1)\,.$$

5.2 $N > 2$

Recall that the equilibrium distribution $G(t)$ for $t \geq 0$ is specified by the differential equation (8), which involves the empty-queue probabilities $P_0(t)(\overset{\triangle}{=} P\{\mathbf{Q}^{-i}(t) = 0\})$. We first present the evolution equations that allow to compute $P_0(t)$ for a symmetric profile $\mathcal{F}^{-i} = G^{N-1}$. These are in fact a simplified version of Lemma 5, which utilizes the symmetry in the arrival distributions. Thus, we consider the system with $M = N-1$ users, and arrival profile $\mathcal{F}^{-i} = G^M$. Let $m(t)$ denote the *number* of users that arrive up to time t (inclusive), and let $p_t(k,m)$ denote the probability that at time t there are k users in the queue and that m users have arrived. Then $(k(t), m(t))$ is a Markov chain, and the flow balance equations yield

$$\begin{aligned}\frac{d}{dt}p_t(k,m) = &-\left(\mu 1_{\{k>0\}} + (M-m)\frac{G'(t)}{1-G(t)}\right)p_t(k,m) \\ &+ 1_{\{k>0\}}(M-m+1)\frac{G'(t)}{1-G(t)}p_t(k-1,m-1) \\ &+ \mu 1_{\{k<m\}}p_t(k+1,m)\end{aligned} \tag{18}$$

for $0 \leq k \leq m \leq M$. As there is no service before $t = 0$, the initial conditions at $t = 0$ are readily seen to be: $p_0(m,m) = \binom{M}{m}G(0)^m(1-G(0))^{M-m}$, and $p_0(k,m) = 0$ if $k \neq m$. Clearly $P_0(t) = \sum_{m=0}^{M} p_t(0,m)$.

To compute the equilibrium distribution G, we may consider $G(0) \in (0,1)$ as a parameter, and compute $G(t)$ for $t > 0$ using the differential equation (8), coupled with (18). These may be integrated up to the first time τ where either $G' = 0$ or $G = 1$. The equilibrium is obtained if both conditions occur simultaneously.

A search procedure is required in order to find the correct value of $G(0)$ that satisfies the boundary conditions. Such a procedure may be efficiently implemented by using the monotonicity properties in Theorem 1; from these, if $G' = 0$ occurs before $G = 1$ then $G(0)$ should be increased, and vice versa.

In Figure 2, we plot the equilibrium density $G'(t)$ that was obtained numerically for several values of N, with system parameters $\alpha = 2$, $\beta = 1$ and $\mu = N\mu_0$ with $\mu_0 = 1$. Note that the

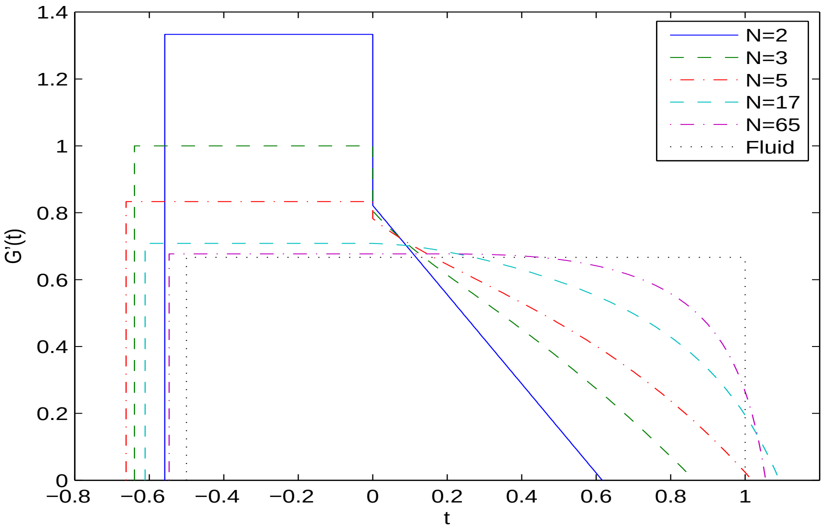

Figure 2: Numerically computed equilibrium arrival densities $G'(t)$ for $\lambda = 2$, $\beta = 1$ and $\mu = N$, comparing several values of N.

service rate is scaled by the number of users, so that the total service-time requirement is kept constant; this convenient for comparison purposes[2]. The fluid solution for this model (with $\mu = 1$) is a uniform distribution on $[-0.5, 1]$, and is depicted in the dotted line. The numerically-obtained result for $N = 2$ coincides with the analytical solution above. For $t < 0$, a uniform density of value $\frac{\mu}{M}\frac{\alpha+\beta}{\beta} = \frac{2}{3}\frac{N}{N-1}$ is obtained, in accordance with Theorem 1(iv). The first arrival time t_a can be seen to be always smaller than its fluid value (see Lemma 10 below), although it is relatively insensitive to the value of N. It may further be seen that both t_a and t_b are not monotone in N. As expected, the normalized densities approach the uniform fluid solution as N is increased.

6 Random Population Size

Up till now it was assumed that N, the number of arrivals to the system, is a deterministic constant. Consequently, each user supposes that there are $M = N - 1$ other arrivals. Here we consider the more general case where the number of arrival may be stochastic,

[2] Exactly the same result is obtained if we keep μ fixed, and then scale the resulting distribution $G(t)$ as $G(Nt)$.

and consequently users only have some probability distribution on M, the number of other arrivals. Let $p_M = (p_M(m), m \geq 0)$ denote this probability distribution, assumed identical for all users, with finite mean $E(M)$.

Let us first clarify the relation between M and N in the stochastic case[3]. Here the simple relation $M = N - 1$ is not valid any more. Following [15], let us derive the distribution of M based the distribution p_N of N. Assume for the moment that N is bounded by a constant N_0. Consider a specific user, say C_1, in this group of N_0 potential arrivals, and let A_1 denote the event that C_1 is an actual arrival. Then, invoking symmetry, we obtain for $n \geq 1$:

$$\begin{aligned} p_M(n-1) = P\{N = n | A_1\} &= \frac{P\{A_1 | N = n\} P\{N = n\}}{P\{A_1\}} \\ &= \frac{\frac{n}{N_0} p_N(n)}{\sum_{n' \geq 1} \frac{n'}{N_0} p_N(n')} = \frac{n p_N(n)}{\sum_{n' \geq 1} n' p_N(n')} \\ &= \frac{1}{E(N)} n p_N(n) \,. \end{aligned} \tag{19}$$

By a limit argument, the same formula holds when N is unbounded. By inverting this relation, p_N can be derived from p_M, with $p_N(0)$ serving as a free parameter. We note that for a binomial distribution, $N \sim B(N_0, p)$, we obtain $M \sim \mathrm{B}(N_0 - 1, p)$, so that $E(M) = E(N) - p$. For the Poisson distribution, $N \sim \mathrm{Pois}(\lambda)$, we interestingly obtain that M has the same distribution, hence $E(M) = E(N)$. These relations can be explained of course by the independence and memoryless properties of the distributions involved. Further discussion on this matter can be found in [7], where it is also pointed out that the Poisson distribution is the only one with this property.

Suppose first that M is bounded by some $M_0 < \infty$. We can essentially repeat our arguments for the deterministic case, to obtain the following result.

Theorem 3 *Suppose M is a bounded random variable, with $P\{M = 0\} < 1$. Then Theorem 1 holds as stated, with M replaced by $E(M)$.*

Proof: (Outline) We will not repeat the detailed arguments, but rather point to some required differences. The arrival profile $\mathcal{F} = (F_i, \, i = 1, \ldots, N_0)$ now contains the arrival time distribution of all potential arrivals. The evolution equations in Lemma 5 still hold separately for each possible set of eventual arrivals, and can be averaged to obtained the

[3] Note that only M is required to determine the equilibrium arrival distribution. N then determines the overall system load.

unconditioned probabilities $p_t(k,n)$. From these the expected queue size $Q(t)$. Repeating the same procedure with user i excluded, we may compute $Q^{-i}(t)$ and $P\{\mathbf{Q}^{-i}(t)=0\}$. Equation (7) for the cost then remains the same. The expected number of arrivals $F(t) = E(\mathbf{A}(t))$ can be similarly computed by averaging over the set of eventual arrivals, and similarly for $F^{-i}(t)$. As a consequence, the required properties of Q^{-i}, $P\{\mathbf{Q}^{-i}(t)=0\}$ and F^{-i} are inherited from the deterministic case, and the proof of symmetry (Proposition 1) follows as before. Focusing on the symmetric case where $F_i \equiv G$, it is easily seen that $F^{-i}(t) = E(M)G(t)$. It therefore follows that Lemma 8 holds with M replaced by $E(M)$, and this extends to all the equilibrium properties in Theorem 1. □

When N is unbounded, the set of evolution equations in Lemma 5 equations is countable, and some care may be required to obtain the continuity properties used in our arguments. While we maintain that the above results carry through, we will not go into the detailed arguments here and postulate this as a conjecture.

We finally address the computation of the symmetric equilibrium for random N case, which requires the solution of the analog of the evolution equations (18). As mentioned above, one option is to compute the respective probabilities conditioned on each possible value of eventual arrivals M, and then average of M to obtain the unconditioned probabilities $p_t(k,m)$. A more efficient option is as follows. Let $R_t(m) = E(M-m|m(t)=m)$ denote the expected number of remaining arrivals, given that m already arrive by time t. By Bayes rule, for $k \geq 0$,

$$\begin{aligned} P(M=m+k|m(t)=m) &= \frac{P(m(t)=m\,|\,M=m+k)P(M=m+k)}{P(m(t)=m)} \\ &= \frac{\binom{m+k}{m}G(t)^m(1-G(t))^k p_M(m+k)}{\sum_{k\geq 0}\binom{m+k}{m}G(t)^m(1-G(t))^k p_M(m+k)} \end{aligned}$$

so that

$$R_t(m) = \sum_{k\geq 0} k\,P(M=m+k\,|\,m(t)=m)\,.$$

Equations (18) can now be seen to hold for $0 \leq k \leq m \leq M_0$, where M_0 is the maximal value that M can attain, after replacing $(M-m)$ by $R_t(m)$, and $(M-m+1)$ by $R_t(m-1)$.

Two special cases are of interest. When $M \sim \mathrm{B}(M_0,p)$ is Binomial, it may be verified that $p(k) \overset{\triangle}{=} P(M=m+k|m(t)=m)$ has $\mathrm{B}(M_0-m, 1-G(t))$ distribution, so that

$R_t(m) = (M_0 - m)(1 - G(t))$. Substituting in (18) and eliminating $(1 - G(t))$ gives

$$\begin{aligned}\frac{d}{dt}p_t(k,m) = &- \left(\mu 1_{\{k>0\}} + (M_0 - m)G'(t)\right) p_t(k,m) \\ &+ 1_{\{k>0\}}(M_0 - m + 1)G'(t)p_t(k-1, m-1) \\ &+ \mu 1_{\{k<m\}} p_t(k+1, m)\,. \end{aligned} \tag{20}$$

Finally, when M is Poisson with mean Λ, we obtain (either through the above formula, or directly using the memoryless property) that $R_t(m) = \Lambda(1 - G(t))$, independent of m. Consequently, by averaging over m we can obtain the following evolution equations in terms of $P_t(k)$ (probability that there are k users in queue at time t) directly:

$$\begin{aligned}\frac{d}{dt}P_t(k) = &- \left(\mu 1_{\{k>0\}} + \Lambda G'(t)\right) P_t(k) \\ &+ 1_{\{k>0\}}\Lambda G'(t)P_t(k-1) + \mu P_t(k+1)\,, \quad k \geq 0\,. \end{aligned} \tag{21}$$

These indeed coincide with the equations that were used in [3] for the Poisson case.

7 Convergence to the Fluid Limit

We next consider the system as the number of users N increases, and show that the equilibrium arrival profile, suitably scaled, converges to the uniform equilibrium profile of the fluid system that was studied in [9, 10].

Recall that in the single-class fluid model, the user population has mass Λ, with each user represented by a point in the interval $[0, \Lambda]$. The cost function is identical to the present paper. Service is deterministic at rate μ, so that the entire user population may be served in $\frac{\Lambda}{\mu}$ time units. Take $\Lambda = 1$ for simplicity. In this setting, the aggregate equilibrium arrival profile is uniform between times $[-\frac{\beta}{\alpha\mu}, \frac{1}{\mu}]$. That is, users arrive at a constant rate of $\mu\frac{\alpha}{\alpha+\beta}$ over that interval. In particular, a fraction $\frac{\beta}{\alpha+\beta}$ of the user population arrives before the opening time at 0.

Let G_N denote the equilibrium arrival profile in the finite population system with N users. We mainly consider the deterministic-N case, and will comment on the stochastic case later in this section. Recall that G_N is obtained as a solution to the differential equation (8), has finite support $[t_a^N, t_b^N]$, and satisfies the boundary conditions $G_N(t_b^N) = 1$ and $G'_N(t_b^N) = 0$. When no confusion arises we will henceforth drop the explicit index N from G, t_a and t_b. Our goal is to show that $G_N(Nt) \equiv G(Nt)$ converges to the uniform fluid profile as

$N \to \infty$, that is, $\frac{d}{dt}G(Nt) \to \mu\frac{\alpha}{\alpha+\beta}$ for $t \in (-\frac{\beta}{\alpha\mu}, \frac{1}{\mu})$. Note that we scale the time axis by N, since the total service time requirement increases proportionally to N. (Alternative we could increase the service rate N-fold.)

The following theorem establishes the required convergence, and provides bounds on the convergence rate.

Theorem 4 *$G(Nt)$ converges to a uniform distribution on $[-\frac{1}{\mu}\frac{\beta}{\alpha}, \frac{1}{\mu}]$ at rate $o(\frac{\log N}{N})$, in the sense that*

$$\frac{N-1}{\mu}G'(t) = \frac{\alpha}{\alpha+\beta} - P_0(t)1_{\{t\geq 0\}}\,, \quad t \in [t_a, t_b]$$

is satisfied with

$$\frac{t_a}{N-1} = -\frac{1}{\mu}\frac{\beta}{\alpha} - o(\frac{\log N}{\sqrt{N}}) < -\frac{1}{\mu}\frac{\beta}{\alpha} \tag{22}$$

$$\frac{t_b}{N-1} = \frac{1}{\mu} + o(\frac{\log N}{\sqrt{N}}) \tag{23}$$

and

$$P_0(t) \leq \frac{1}{N} \quad \textit{for} \quad \frac{t}{N} \leq \frac{1}{\mu} - o(\frac{\log N}{\sqrt{N}})\,. \tag{24}$$

We proceed to prove this result. We will find it convenient to state all claims in terms of of $M = N-1$ as the index rather than N itself; this of course has no effect on the stated rates. Given the characterization of the equilibrium in Theorem 1, the key to this convergence result is in showing that the empty-queue probability $P_0(t) = P\{\mathbf{Q}^{-i}(t) = 0\}$ is small as long as t is not too close to $\frac{M}{\mu}$. For that to hold, however, we need to show that enough users choose to arrive early enough. We start by providing bounds on the support $[t_a, t_b]$ of the equilibrium distribution G.

Lemma 10 *For all $M \geq 1$,*

$$G(0) = \frac{\beta}{\alpha+\beta} + d_M \quad \textit{for some } d_M > 0\,, \tag{25}$$

and consequently

$$t_a = -\frac{M}{\mu}\frac{\beta}{\alpha}(1 + \frac{\alpha+\beta}{\beta}d_M)\,. \tag{26}$$

Furthermore,

$$\frac{M}{\mu}(1 - \frac{\alpha+\beta}{\alpha}d_M) < t_b \leq \frac{M}{\mu}(1 + \frac{\alpha+\beta}{\beta}d_M)\,. \tag{27}$$

Proof: We first show that $d_M > 0$. Suppose to the contrary that $d_M \leq 0$, or $G(0) \leq \frac{\beta}{\alpha+\beta}$. Note that an arrival at $t = 0$ sees an expected queue length of $MG(0)$ and hence incurs an expected cost of $C_i(0) = (\alpha+\beta)\frac{MG(0)}{\mu}$. Thus, $G(0) \leq \frac{\beta}{\alpha+\beta}$ implies $C_i(0) \leq \frac{M}{\mu}\beta$. Further, $G(0) \leq \frac{\beta}{\alpha+\beta}$ together with (8) and $P_0(t) > 0$ imply that $G(\frac{M}{\mu}) < 1$. Hence $t^* \triangleq \frac{M}{\mu}$ is in the support of G, implying that $C_i(t^*) = C_i(0)$ by the equilibrium property. On the other hand, by (6), $C_i(t^*) = \beta t^* + (\alpha+\beta)Q^{-i}(t^*)$. Since there will be a queue at time $\frac{M}{\mu}$ with positive probability, we obtain $C_i(t^*) > \beta\frac{M}{\mu} \geq C_i(0)$, providing the desired contradiction.

Equation (26) follows by noting that $G'(t) = \frac{\mu}{M}\frac{\alpha}{\alpha+\beta}$ for $t < 0$, so that $t_a = -G(0)\frac{M}{\mu}\frac{\alpha+\beta}{\alpha}$.

To establish (27), note first that at $t' = \frac{M}{\mu}(1 - \frac{\alpha+\beta}{\alpha}d_M)$ we obtain

$$G(t') < G(0) + \frac{\mu}{M}\frac{\alpha}{\alpha+\beta}t' = \frac{\beta}{\alpha+\beta} + d_M + \frac{\alpha}{\alpha+\beta}(1 - \frac{\alpha+\beta}{\alpha}d_M) = 1\,,$$

so that $t' < t_b$. To upper-bound t_b, note again that $C_i(0) = (\alpha+\beta)\frac{MG(0)}{\mu}$. However an arrival at time $t > 0$ incurs a cost of $C_i(t) > \beta t$, and $C_i(t_b) = C_i(0)$ leads to the rightmost inequality in (27). □

The following upper bound on the empty-queue probability is obtained by applying Chernoff's bound to the arrival and service processes.

Lemma 11 *Consider $t \geq 0$ so that $MG(t) > \mu t$. Then*

$$P_0(t) \triangleq P\{\mathbf{Q}^{-i}(t) = 0\} \leq \exp\left(-\frac{1}{2}\frac{(MG(t)-\mu t)^2}{MG(t)+\mu t}\right)\,. \tag{28}$$

Proof: Fixing t, let $\mathbf{A}$ denote the total number of arrivals by t, and $\mathbf{D}$ the total number of *potential* service completions (assuming no idleness) by that time. Further denote $p = G(t)$. Then $\mathbf{A} \sim B(M,p))$, a binomial random variable, and $\mathbf{D} \sim \text{Pois}(\mu t)$, a Poisson random variable independent of $\mathbf{A}$. It is evident that $\mathbf{Q}^{-i} \geq \mathbf{A} - \mathbf{D}$, and therefore

$$P_0(t) \leq P\{\mathbf{A} - \mathbf{D} \leq 0\}\,.$$

Note that $E(\mathbf{A} - \mathbf{D}) = Mp - \mu t > 0$ by assumption. Now, for any $v > 0$,

$$\begin{aligned} P\{\mathbf{A} - \mathbf{D} \leq 0\} &\leq E(e^{-v(\mathbf{A}-\mathbf{D})}) = E(e^{-v\mathbf{A}})E(e^{v\mathbf{D}}) \\ &= (1 - p(1-e^{-v}))^M \exp(\mu t(e^v - 1)) \\ &= \exp\left(M\log(1 - p(1-e^{-v})) + \mu t(e^v - 1)\right) \\ &\leq \exp\left(-Mp(1-e^{-v}) + \mu t(e^v - 1)\right)\,, \end{aligned}$$

as $\log(1-x) < -x$. Choosing $e^v = \sqrt{\frac{Mp}{\mu t}}$ we obtain

$$P\{\mathbf{A} - \mathbf{D} \leq 0\} \leq \exp(-(\sqrt{Mp} - \sqrt{\mu t})^2)\,.$$

Noting that $(\sqrt{a} - \sqrt{b})^2 = \frac{(a-b)^2}{(\sqrt{a}+\sqrt{b})^2} \geq \frac{(a-b)^2}{2(a+b)}$, the required bound is established. □

Proof of Theorem 4: Observing equation (8), we have

$$G(t) = G(0) + \frac{\mu}{M}(\frac{\alpha}{\alpha+\beta}t - \int_0^t P_0(s)ds)\,. \tag{29}$$

But $G(0) > \frac{\beta}{\alpha+\beta}$ by (25), so that

$$G(t) > \frac{\beta}{\alpha+\beta} + \frac{\mu}{M}(\frac{\alpha}{\alpha+\beta}t - \int_0^t P_0(s)ds)\,.$$

We can now iterate this inequality together with (28) to upper-bound $P_0(t)$. Taking some $0 < Z_M < M\frac{\alpha}{\alpha+\beta}$, we will show that $MG(t) - \mu t \geq Z_M$ holds up to some $t_3 < \frac{M}{\mu}$. For $t = 0$ this is true since $MG(0) > M\frac{\alpha}{\alpha+\beta} > Z_M$. Now, if $MG(t) - \mu t > Z_M$ for $t \leq t_3 < \frac{M}{\mu}$, then

$$P_0(t) \leq \exp\left(-\frac{Z_M^2}{4M}\right) \triangleq p_M\,, \tag{30}$$

where we have used the fact that $MG(t) + \mu t \leq 2M$ for $t < \frac{M}{\mu}$. Therefore,

$$G(t) \geq \frac{\beta}{\alpha+\beta} + \frac{\mu}{M}(\frac{\alpha}{\alpha+\beta}t - p_M t) \geq \frac{\beta}{\alpha+\beta} + \frac{\mu}{M}\frac{\alpha}{\alpha+\beta}t - p_M\,,$$

and

$$MG(t) - \mu t \geq M\frac{\beta}{\alpha+\beta} - \mu\frac{\beta}{\alpha+\beta}t - Mp_M\,. \tag{31}$$

Inequalities (30) and (31) remain valid as as long as $MG(t) - \mu t \geq Z_M$, which by (31) is guaranteed for all

$$t \leq \frac{M}{\mu} - \frac{M}{\mu}\frac{\alpha+\beta}{\beta}(p_M + \frac{Z_M}{M}) \triangleq t_3\,.$$

Choose now $Z_M = \sqrt{4M\log M}$, so that $p_M = M^{-1}$. Observe that

$$t_3 = \frac{M}{\mu} - \frac{M}{\mu}\frac{\alpha+\beta}{\beta}(\frac{1}{M} + \frac{\sqrt{4M\log M}}{M}) = \frac{M}{\mu}(1 - o\left(\frac{\log M}{\sqrt{M}}\right))\,.$$

We have thus established (24). To show (22)-(23), observe that $G'(t_3) = \frac{\mu}{M}(\frac{\alpha}{\alpha+\beta} - P_0(t_3)) \geq \frac{\mu}{M}(\frac{\alpha}{\alpha+\beta} - \frac{1}{M}) > 0$ for M large enough, so that $t_3 < t_b$, hence $G(t_3) < 1$. But using the derived bound $P_0(t) \leq p_M$ in (29) gives

$$G(t_3) \geq G(0) + \frac{\mu}{M}(\frac{\alpha}{\alpha+\beta} - p_M)t_3 \geq G(0) + \frac{\mu}{M}\frac{\alpha}{\alpha+\beta}t_3 - p_M\,,$$

(observe that $t_3 < \frac{M}{\mu}$ by its definition), so that $G(t_3) < 1$ implies

$$\begin{aligned} G(0) &< 1 - \frac{\mu}{M}\frac{\alpha}{\alpha+\beta} t_3 + p_M \\ &= \frac{\beta}{\alpha+\beta} + (\frac{1}{M} + \frac{\sqrt{4M\log M}}{M})\frac{\alpha}{\beta} + \frac{1}{M} \,. \end{aligned}$$

Therefore

$$d_M \triangleq G(0) - \frac{\beta}{\alpha+\beta} = o\left(\frac{\log M}{\sqrt{M}}\right) \,,$$

and (22)-(23) follow by Lemma 10. □

Remark 2 *The bound on $P_0(t)$ in* (24) *actually becomes much tighter as we decrease t and move away from the boundary of the support. Indeed, using the notation of the last proof, for $t < t_3$ we have*

$$MG(t) - \mu t \geq \frac{\beta}{\alpha+\beta}(t_3 - t) + MG(t_3) - \mu t_3 = \frac{\beta}{\alpha+\beta}(t_3 - t) + Z_M \,,$$

which can be used in (30) *in lieu of Z_M.*

We finally comment on the model with a random number of arrivals, as presented in Section 6. Suppose that (M_i) is a sequence of random variables, with $E(M_i) \to \infty$. Suppose further that $\frac{M_i}{E(M_i)}$ convergence to a deterministic constant (without this assumption the fluid limit, if such exists, could be quite different from the one above). In that case, essentially the same proof can be applied to show convergence to the fluid limit above; naturally the bound in Lemma 11 may be different, and lead to different convergence rates. However, in the important cases where the M's are Binomially distribution with fixed success probability p, or Poisson distributed, it is readily verified that the same bound holds with M replaced by its mean $E(M)$. Therefore, the convergence results above should hold as stated, save for this substitution.

8 Price of Anarchy

The price of anarchy (PoA) is a well accepted measure for the social inefficiency of the non-cooperative equilibrium solution. It corresponds to the ratio of the social cost in the worst Nash equilibrium, to the cost of the socially optimal solution, obtained here when the arrival times are optimally determined by a central planner.

The socially optimal solution naturally depends on the restrictions imposed on the central planner. One option in our case is to assume that the arrival times must be pre-scheduled, with no on-line feedback on service completions. While this may pose a reasonable choice in our setting, the explicit solution of the the resulting optimization problem appear to be hard (except in the simple case of $N = 2$)[4], and will not be pursued here. The second option is to allow the controller to schedule arrivals on-line based on observed service completions[5]. In that case the solution is obvious: users are scheduled one after another, starting at time 0, with no waiting or idleness. It is easy to see that the social cost in this case (the sum over the expected user costs of the N users) is given by

$$W^* = \frac{\beta}{\mu}(1 + 2 + \cdot + (N-1)) = \frac{\beta}{\mu}\frac{N(N-1)}{2} .$$

This will be used here for the baseline, socially optimal solution. Recall that in the homogeneous fluid model studied in [10], the price of anarchy is exactly 2. We then have the following result for our finite-user model.

Proposition 5 *PoA > 2 for all $N \geq 2$, and it converges to 2 as $N \to \infty$, at rate $o(\frac{\log N}{\sqrt{N}})$.*

Proof: Recall that $t_a < 0$ is the first point in the support of the equilibrium distribution $G(t)$. As an arrival at t_a would be served first, the equilibrium cost for each user is $-\alpha t_a$, and the PoA turns out to be

$$\text{PoA} = \frac{N\alpha(-t_a)}{W^*} = 2\frac{\alpha\mu(-t_a)}{\beta(N-1)} .$$

The bound on t_a in equation (22) of Theorem 4 (where $M = N - 1$) now imply the stated results. □

We finally note that for $N = 2$, the explicit solution obtained in Section 5.1 yields PoA $= 2\sqrt{1 + 2\alpha/\beta}$.

9 Conclusion

We have addressed here the strategic choice of arrival times into a transient FCFS queue, where each user balances the benefit of early service completion with that of a short wait in

[4] The optimal social cost in the case of $N = 2$ turns out to be $\mu^{-1}\beta\left(1 + \log\left(\frac{\alpha+\beta}{\beta}\right)\right)$, with the first user arriving at $t = 0$ and the second at $t = \mu^{-1}\log\left(\frac{\alpha+\beta}{\beta}\right)$. The PoA can still be seen to be larger than 2 relative to this solution.

[5] These two options coincide in the fluid model

the queue. Distinct contributions include establishing the existence and uniqueness of the equilibrium without assuming a-priori that the equilibrium is symmetric; the consideration of a general number of arrivals, not necessarily of Poisson distribution; demonstrating convergence to the fluid limit; and identifying the asymptotic price of anarchy for this model.

The model we considered does not specify a closing time for the server, and allows queueing before the opening time. However, the essential approach and results should be easily extendable under the opposite assumptions, similarly to [3, 6, 7]. Another variant of interest to the current model pertains to the cost function: Suppose that instead of caring for early service completions, the users are interested to get served before others. This may be the case, for example, in a box-office queue, where the first customers get the better seats. The two cost variants coincide in the fluid model (as noted in [10]), in the finite-population model this requires separate consideration, which is left to future work.

Our model may be further extended in several obvious and important directions. One is the extension to the multi-class model, with a non-homogeneous user population. Another is the consideration of non-linear cost functions – in particular, it will be interesting to replace the linear tardiness cost with a V-shaped (dis)-punctuality cost, as is common in the transportation bottleneck model. The analysis of the equilibrium in these extended models presents a considerable challenge for future work.

Let us conclude the paper with a word on the significance of the equilibrium solution. As we have shown, the unique equilibrium here prescribes for each user a probability distribution over a continuous interval, in which all points identical costs. This is of course the usual case with a (mixed) Nash equilibrium, and much has been said about the interpretation of this equilibrium as a descriptive (rather than prescriptive) solution. In the context of our model, it should be realized that we do not necessarily expect each user to fully randomize his choice. Rather, the equilibrium reflects each user beliefs about the others' choices, as finally reflected by the queue he expects to see upon his arrival. These beliefs may be the result of general prior experience (as, for example, in the case of one-time queue forming for a specific gadget) – or alternatively, the result of a repeated interaction and learning with the specific queueing system (as, for example, in the case of queueing in the cafeteria at work, or selecting the start time for commuting to work). Proposing a specific learning mechanism and studying its convergence properties vis-a-vis the equilibrium solution is again a challenging task for future work.

Acknowledgments

The first author would like to thanks Yahoo Research India Lab for partially funding this research. This research was additionally supported by the Israel Science Foundation, grant No. 1319/11.

References

[1] G. Allon and I. Gurvich, "Pricing and dimensioning competing large-scale service providers", *Manufacturing Service Oper. Management* 12(3):449-469, 2010.

[2] M. Armony and C. Maglaras, "On customer contact centers with a call-back option: customer decisions, routing rules, and system design", *Oper. Res.* 52(2):271-292, 2004.

[3] A. Glazer and R. Hassin, "?/M/1: On the equilibrum distribution of user arrivals", *Eur. J. Oper. Res.* 13:146-150, 1983.

[4] J.K. Hale and S.M Verduyn Lunel, *Introduction to Functional Differential Equations,* Springer, 1993.

[5] R. Hassin and M. Haviv, *To Queue or Not to Queue*, Kluwer Academic Publishers, 2003.

[6] R. Hassin and Y. Kleiner, "Equilibrium and optimal arrival patterns to a server with opening and closing times", *IEE Transactions* 43(3):164-175, March 2011.

[7] M. Haviv, "When to arrive at a queue with tardiness costs?", preprint, October 2010.

[8] H. Honnappa and R. Jain, "Strategic arrivals into queueing networks", Proc. 48th Annual Allerton Conference, Illinois, Oct. 2010, pp. 820-827.

[9] S. Juneja and R. Jain, "The Concert/Cafeteria Queuing Problem: A Game of Arrivals", in *Proc. ValueTools'09 – Fourth ICST/ACM Fourth International Conference on Performance Evaluation Methodologies and Tools,* Pisa, Italy, October 2009.

[10] R. Jain, S. Juneja and N. Shimkin, "The Concert Queuing problem: To wait or to be late", *Discrete Event Dyn. Syst.* 21:103-138, 2011.

[11] S. Kumar and R.S. Randhawa, "Exploiting market size in service systems", *Manufacturing Service Oper. Management* 12(3):511-526, 2010.

[12] J.-M. Lasry and P.-L. Lions, "Mean field games". *Jpn. J. Math.* 2(1):229260, 2007.

[13] R. Lindley, "Existence, uniqueness, and trip cost function properties of user equilibrium in the bottleneck model with multiple user classes", *Transport. Sci.* 38(3):293-314, 2004.

[14] C. Maglaras and A. Zeevi, "Pricing and design of differentiated services: Approximate analysis and structural insights", *Oper. Res.* 53(2):242262, 2005.

[15] R. P. McAfee and J. McMillan, "Auctions with a stochastic number of bidders", *J. Economic Theory* 43(1):1-19, 1987.

[16] G.F. Newell, "The morning commute for nonidentical travellers", *Transport. Sci.* 21(2):74-88, May 1987.

[17] H. Otsubo and A. Rapoport, "Vickrey's model of traffic congestion discretized", *Transport. Res. B* 42:873-889, 2008.

[18] A. Rapoport, W.E. Stein, J.E. Parco and D.A. Seale, "Equilibrium Play in Single-Server Queues with Endogenously Determined Arrival Times", *Journal of Economic Behavior and Organization* 55:67-91, 2004.

[19] D. Seale, J. Parco, W. Stein and A. Rapoport, "Joining a Queue or Staying Out: Effects of Information Structure and Service Time on Arrival and Staying Out Decisions", *Experimental Economics* 8(2):117-144, June 2005.

[20] W.S. Vickrey, "Congestion theory and transport investment", *American Econ. Rev.* 59:251-260, 1969.

Appendix

A Proofs for Section 3

Proof of Lemma 3: We use stochastic coupling with the two arrival process implemented on a common probability space, and with identical service times for all users. Noting that $\mathbf{A}(s_2) - \mathbf{A}(s_1) = \sum_i 1_{\{s_1 < T_i \leq s_2\}}$ for $s_1 < s_2$, where $T_i \sim F_i$, it follows by the assumed dominance relation in (i) that $\mathbf{A}(s_2) - \mathbf{A}(s_1) \geq \tilde{\mathbf{A}}(s_2) - \tilde{\mathbf{A}}(s_1)$ for all $s_1 < s_2 \leq t$. That is, the number of arrivals under $\mathcal{F} = \{F_i\}$ is at least as large as under $\tilde{\mathcal{F}}$, over *any* time span up to t. This clearly implies that $\mathbf{Q}(t) \geq \tilde{\mathbf{Q}}(t)$ w.p. 1, hence $Q(t) \geq \tilde{Q}(t)$. As for (ii), with the same coupling it now follows that $\mathbf{A}(t) > \tilde{\mathbf{A}}(t)$ with positive probability, hence $\mathbf{Q}(t) > \tilde{\mathbf{Q}}(t)$ on an event of positive probability (e.g., when there are no service completions by time t), so that $Q(t) > \tilde{Q}(t)$. Assertion (iii) follows by similar considerations applied to the indicator $1_{\{\mathbf{Q}(t)>0\}}$ in place of $\mathbf{Q}(t)$. □

Proof of Lemma 4: It is sufficient to establish the claim for the case where the arrival profiles are the same except for one component, as it then extends to the general case via the triangle inequality. Suppose then that $F_i = \tilde{F}_i$ for $i \geq 2$, we need to show that

$$|P\{\mathbf{Q}(t) = 0\} - P\{\tilde{\mathbf{Q}}(t) = 0\}| \leq K\|F_1 - \tilde{F}_1\|_t \,.$$

Suffices to show that $P\{\mathbf{Q}(t) > 0\} = 1 - P\{\mathbf{Q}(t) = 0\}$ is Lipschitz continuous. In particular, we argue that

$$|P\{\mathbf{Q}(t) > 0\} - P\{\tilde{\mathbf{Q}}(t) > 0\}| \leq 2\|F_1 - \tilde{F}_1\|_t. \tag{32}$$

To see this, note that (32) holds whenever, $\|F_1 - \tilde{F}_1\|_t > \frac{1}{2}$, since,

$$|P\{\mathbf{Q}(t) > 0\} - P\{\tilde{\mathbf{Q}}(t) > 0\}| \leq 1.$$

Now suppose that $(\tilde{F}_1(s) : s \leq t)$ satisfies the constraint $\|F_1 - \tilde{F}_1\|_t \leq \epsilon \leq \frac{1}{2}$. Then, it maximizes $P\{\tilde{\mathbf{Q}}(t) > 0\}$ when

$$\tilde{F}_1(s) = \max(0, F_1(s) - \epsilon)$$

for $0 \leq s < t$ and $\tilde{F}_1(t) = \min(1, F_1(t) + \epsilon)$. To see this, consider any distribution function H such that $\|F_1 - H\|_t \leq \epsilon$. We can couple an arrival from H and $\tilde{F}_1$ using the same uniform random variable U distributed uniformly over $[0, 1]$.

Define for any distribution function $R(\cdot)$

$$R^{-1}(u) = \inf\{x : R(x) \geq u\}.$$

Then, since $H(s) \geq \tilde{F}_1(s)$ for $s < t$, it follows that for $U < H(t^-)$, $H^{-1}(U) < t$ and

$$H^{-1}(U) \leq \tilde{F}_1^{-1}(U).$$

Furthermore, $U \leq H(t)$ is equivalent to $H^{-1}(U) \leq t$. It then follows that

$$U \leq H(t) \leq F_1(t) + \epsilon$$

so that then $\tilde{F}_1^{-1}(U) \leq t$. Hence, $U \leq H(t)$ implies $H^{-1}(U) \leq \tilde{F}_1^{-1}(U) \leq t$, so that

$$P\{\tilde{\mathbf{Q}}(t) > 0\} \geq P\{\mathbf{Q}_H(t) > 0\},$$

where $P\{\mathbf{Q}_H(t) > 0\}$ corresponds to $P\{\mathbf{Q}(t) > 0\}$ with distribution F_1 replaced by H. Now it is easy to see that

$$P\{\tilde{\mathbf{Q}}(t) > 0\} - P\{\mathbf{Q}(t) > 0\} \leq 2\epsilon.$$

To see this, let U again denote uniform random variable as above. Let $V = U - \epsilon$ for $U \geq \epsilon$ and $V = 1 - U$ for $U < \epsilon$. It is easy to see that V is also uniformly distributed over $[0, 1]$.

We develop a stochastic coupling where U is used to generate samples from F_1 and V is used to develop samples from $\tilde{F}_1$. Samples from the remaining distributions $(F_i : 2 \leq i \leq N)$ are kept the same in the two systems.

Then note that $F_1^{-1}(U) = \tilde{F}_1^{-1}(V)$ except if $U \leq \epsilon$ or $U \geq F_1(t)$. For $U > F_1(t) + \epsilon$, the arrivals occur after t under both the distributions. Hence, under the two systems, the arrivals that occur before or at time t have different times with at most 2ϵ probability. Now, (32) easily follows. □

B Proof of Proposition 1: Symmetry

The proof proceeds through several lemmas. The first establishes that if a certain user assigns a positive probability of arrival to a single point in time, then other users will choose not to arrive at or shortly after that time.

Lemma 12 *Let $\mathcal{F}$ be an equilibrium profile. Suppose F_i has a point mass at t, namely, $F_i(t) > F_i(t-)$. Then* *(i)* $F_j(t+\epsilon) - F_j(t-) = 0$ *for some* $\epsilon > 0$ *and all* $j \neq i$.

(ii) In particular, F_j *does not have a point mass at* t.

Proof: Claim (ii) clearly follows from (i). To establish (i), we show that if user $j \neq i$ arrives just before t, he will incur a cost that is smaller than if he arrives at t or shortly thereafter. Therefore, arriving at t or shortly thereafter is not a valid choice at equilibrium.

The expected queue size faced by user j arriving at t is

$$\begin{aligned}\bar{Q}^{-j}(t) &= \frac{1}{2}(Q^{-j}(t-) + Q^{-j}(t)) \\ &= Q^{-j}(t-) + \frac{1}{2}(F^{-j}(t) - F^{-j}(t-)) \\ &\geq Q^{-j}(t-) + \frac{1}{2}(F_i(t) - F_i(t-)) \; > \; Q^{-j}(t-)\,.\end{aligned}$$

Here the second equality follows by Lemma 2, the first inequality holds since $F^{-j} = \sum_{k \neq j} F_k$ includes F_i as an additive term, and the last inequality holds since F_i does have a point mass at t by assumption. Note that $Q^{-j}(t-\epsilon) \to Q^{-j}(t-)$ as $\epsilon \downarrow 0$. Observing (7), it follows that $C_j(t-\epsilon_1, \mathcal{F}^{-j}) < C_j(t, \mathcal{F}^{-j})$ for ϵ_1 small enough.

Further, since the cost is right-continuous (Lemma 7), this inequality extends to $s > t$, namely $C_j(t-\epsilon_1, \mathcal{F}^{-j}) < C_j(s, \mathcal{F}^{-j})$ for $s \in [t, t+\epsilon]$ with $\epsilon > 0$ small enough. This means that arriving at any point in $[t, t+\epsilon_1]$ is not an optimal choice for j, so that F_j must assign zero probability to that interval. Hence $F_j(t+\epsilon) - F_j(t-) = 0$, as claimed. □

Recall from Lemma 1 that $C_i(t, \mathcal{F}^{-i}) = c_i$ on a set of F_i-measure 1. We wish to establish that this property holds pointwise on $\mathcal{T}_i$. The next lemma establishes this claim, with the possible exception of points to which other users assign point masses. Later we will show that such point masses do not exist.

Lemma 13 *Let* $\mathcal{F} = \{F_i\}$ *be an equilibrium profile. Then, for each* i,

(i) The support $\mathcal{T}_i$ *of* F_i *is bounded.*

(ii) $C_i(t, \mathcal{F}^{-i}) = c_i$ *if both* $t \in \mathcal{T}_i$ *and* F^{-i} *has no point mass at* t.

(iii) $C_i(t-, \mathcal{F}^{-i}) = c_i$ *for all* $t \in \mathcal{T}_i$.

Proof: (i) Follows since the cost $C_i(s, \mathcal{F}^{-i})$ tends to infinity at $|s| \to \infty$.

(ii) $t \in \mathcal{T}_i$ implies that $(t-\epsilon, t+\epsilon)$ has positive F_i measure for any $\epsilon > 0$. It follows from Lemma 1(ii) that $C_i(s, \mathcal{F}^{-i}) = c_i$ for some point $s \in (t-\epsilon, t+\epsilon)$. But F^{-i} does not have a point mass at t by assumption, so that by Lemma 7, t is a continuity point of $C_i(\cdot, \mathcal{F}^{-i})$. It follows that $C_i(t, \mathcal{F}^{-i}) = c_i$ at t as well.

(iii) Consider again $t \in \mathcal{T}_i$. If F^{-i} has no point mass at t, then t is a continuity point of $C_i(t, \mathcal{F}^{-i})$ by Lemma 7, and the conclusion follows from (ii). Suppose F^{-i} does contain a point mass at t. Lemma 12 then implies that F_i assigns zero probability to $[t, t+\epsilon)$ for some $\epsilon > 0$. But, as argued in (ii), $(t-\epsilon, t+\epsilon)$ has positive F_i measure for any $\epsilon > 0$. It follows that $(t-\epsilon, t)$ has positive F_i measure for all $\epsilon > 0$. It now follows by Lemma 1(ii) that $C_i(s, \mathcal{F}^{-i}) = c_i$ for some point $s \in (t-\epsilon, t)$, for all $\epsilon > 0$. But since the left limit $C_i(t-, \mathcal{F}^{-i})$ exists by Lemma 7, the claim follows. □

We next show that the equilibrium costs are the same for all users. From this, we will subsequently infer that their arrival distribution must be identical as well.

Lemma 14 *In any equilibrium profile $\mathcal{F}$, the equilibrium costs c_i of the users are all the same.*

Proof: It suffices to show that $c_j \le c_i$ for all $j \ne i$. Fix i and j. Let t_i be the smallest point in the support $\mathcal{T}_i$ of F_i (t_i exists since the support is bounded by Lemma 13, and closed by definition). In the next paragraph we will show that $C_j(t_i-, \mathcal{F}^{-j}) \le c_i$; that is, an arrival of j just before t_i will incur a cost not exceeding c_i. (Note that this also applies to an arrival at $t = t_i$ itself, unless F_j has a point mass there.) This inequality clearly implies that $c_j \le c_i$, since c_j minimizes $C_j(t, \mathcal{F}^{-j})$.

By definition of t_i we have that $F_i(t) = 0$ for $t < t_i$, hence $F^{-j}(t) \le F^{-i}(t)$ there. Invoking the monotonicity property in Lemma 3(i), it follows that $Q^{-j}(t_i-) \le Q^{-i}(t_i-)$. Observing (7), this implies that $C_j(t_i-, \mathcal{F}^{-j}) \le C_i(t_i-, \mathcal{F}^{-i})$. But the latter cost equals c_i by Lemma 13(iii). Therefore $C_j(t_i-, \mathcal{F}^{-j}) \le c_i$, and the lemma is established. □

Lemma 15 *Any equilibrium profile $\mathcal{F}$ is symmetric, in the sense that F_i does not depend on i.*

Proof: Suppose, in contradiction, $F_i \ne F_j$ for some i and j. Let $t_0 = \max\{t : F_i(s) = F_j(s)\,,\ s < t\}$ be the maximal time up to which these distributions are identical. Note that $t_0 > -\infty$ since the supports of F_i and F_j are bounded, and it may then be easily seen from the definition that the maximum is indeed attained at a finite point. We consider separately the following two cases:

(i) Either $(t_0, t_0+\epsilon) \not\subset \mathcal{T}_i$ for all $\epsilon > 0$, or $(t_0, t_0+\epsilon) \not\subset \mathcal{T}_j$ for all $\epsilon > 0$. That is, the support of F_i, or that of F_j, does not extend continuously beyond t_0.
(ii) $(t_0, t_0+\epsilon) \subset \mathcal{T}_i \cap \mathcal{T}_j$ for some $\epsilon > 0$. That is, both supports extend to some interval beyond t_0.

Case (i): Suppose the stated condition holds for i (or otherwise swap indices). Observing Assumption 1, there must exist a whole interval $(t_0, t_0+\epsilon)$ which is outside of $\mathcal{T}_i$. Let $t_1 > t_0$ be the first point in $\mathcal{T}_i$ beyond t_0 (such a point must exist since $F_i(t_0) = F_j(t_0) < 1$ since $F_i \neq F_j$ and by definition of t_0). Then $F_i(t_1-) - F_i(t_0) = 0$. But this implies that $F_j(t_1-) - F_j(t_0) > 0$, by definition of t_0. It therefore follows that F_j strictly dominates F_i on $(-\infty, t_1)$ in the sense of Lemma 2, namely $F_j(t) - F_j(s) \geq F_i(t) - F_j(s)$ for all $s < t < t_1$, with strict inequality holding for some $s < t < t_1$. Applying Lemma 2(ii) with $\mathcal{F} = \mathcal{F}^{-i}$ and $\tilde{\mathcal{F}} = \mathcal{F}^{-j}$, we obtain that $Q^{-i}(t_1-) > Q^{-j}(t_1-)$. We next show that this implies $c_i > c_j$, contradicting Lemma 14. Indeed, since $t_1 \in \mathcal{T}_i$, Lemma 13(iii) implies that $C_i(t_1-, \mathcal{F}^{-i}) = c_i$, while the definition of the equilibrium costs (see Lemma 1) implies that $c_j \leq C_j(t_1-, \mathcal{F}^{-j})$. We therefore obtain that $c_i > c_j$, in contradiction to Lemma 14.

Case (ii): Here we shall argue that $F_i(t) = F_j(t)$ for $t \in (t_0, t_0+\epsilon)$. As this stands at odds with the definition of t_0, the required contradiction will be established.[6]

Let k stand for i or j. Since $(t_0, t_0+\epsilon) \subset \mathcal{T}_k$, it follows by Lemma 12(ii) that $F^{-k}(t)$ has no point masses for $t \in (t_0, t_0+\epsilon)$. Therefore, by Lemma 13(ii), $C_k(t, \mathcal{F}^{-k}) = c_k$ on that interval. But $c_i = c_j = c_0$ by Lemma 14, so that $C_i(t, \mathcal{F}^{-i}) = C_j(t, \mathcal{F}^{-j}) \equiv c_0$ for $t \in (t_0, t_0+\epsilon)$. We proceed to show that this implies $F_i(t) = F_j(t)$ for $t \in (t_0, t_0+\epsilon)$.

Consider henceforth $k \in \{i, j\}$ and $t \in (t_0, t_0+\epsilon)$. As F^{-k} has no point masses there we have by Lemma 2 that $Q^{-k}(t)$ is continuous, so that the equality in (6) is in effect. Using the relations in (4) and (3), we obtain

$$\begin{aligned} C_k(t, \mathcal{F}^{-k}) &= (\alpha+\beta)[\mu^{-1}F^{-k}(t) + E(\mathbf{I}^{-k}(t))] - \alpha t \\ &= (\alpha+\beta)[\mu^{-1}F^{-k}(t) + 1_{\{t\geq 0\}} \int_0^t P\{\mathbf{Q}^{-k}(s) = 0\} ds] - \alpha t \\ &\equiv c_0 \,. \end{aligned} \tag{33}$$

[6] It is relatively easy to rule out the case where F_i strongly dominates F_j (or vice versa) over $(t_0, t_0+\epsilon)$, proceeding similarly to case (i). However, such dominance is not entailed in general from $F_i \neq F_j$. For example, suppose $F_i(t) - F_j(t) = t^3 \sin(1/t)$ for $t > 0 \overset{\triangle}{=} t_1$. This is a continuous function that has an infinite number of sign changes near 0, hence neither F_i or F_j dominates the other. We therefore resort to a more elaborate argument involving uniqueness of solutions to a certain differential equation.

Suppose first that $t_0 < 0$. We directly obtain from the last equality that

$$F^{-k}(t) = \frac{\mu}{\alpha+\beta}(\alpha t + c_0)\,, \quad t_0 < t < \min\{0, t_0+\epsilon\}\,. \tag{34}$$

Observe that $F^{-i} = F_0 + F_j$ and $F^{-j} = F_0 + F_i$, where $F_0 \triangleq \sum_{m\neq i,j} F_m$ is common to both. It therefore follows from (34) that $F_j(t) = F_k(t)$ on that interval, which contradicts the definition of t_0.

We may therefore restrict attention to $t_0 \geq 0$. Taking the derivative in (33) and rearranging, we obtain

$$\frac{d}{dt}F^{-k}(t) = \mu\frac{\alpha}{\alpha+\beta} - \mu P\{\mathbf{Q}^{-k}(t) = 0\}\,, \quad k = i, j \tag{35}$$

wherever the derivative exists. But since F^{-k} has no point masses in $[t_0, t_0+\epsilon)$ (as already observed), the right-hand side may be seen to be continuous in t, so that the derivative exists for all $t > 0$ in that range. A possible exception is for $t_0 = 0$ (due to the removal of the indicator), where we simply consider the right derivative.

We next interpret (35) as a coupled pair of functional differential equations for F^{-i} and F^{-j} over $t \geq t_0$. Let us start with the initial conditions $F^{-k}(t_0)$. As we deal with a specific equilibrium profile $\mathcal{F}$, we consider $\{F_m,\ m \neq i,j\}$ as given and fixed. We further consider F_i and F_j as given up to $t < t_0$ (with $F_i = F_j$ there). Since there is no point mass at t_0, then by continuity $F_i(t_0) = F_j(t_0)$ are given as well. Hence $F^{-i}(t_0) = F^{-j}(t_0)$ are also given and serve as initial conditions for (35).

Observe next that $P\{\mathbf{Q}^{-i}(t) = 0\}$ generally depends on $\mathcal{F}^{-i} = \{F_m\,,\ m \neq i\}$. Since we consider $\{F_m,\ m \neq i,j\}$ as given, then $P\{\mathbf{Q}^{-i}(t) = 0\}$ is effectively a function of F_j only, hence of $F^{-i} = F_0 + F_j$. Furthermore, we claim that $P\{\mathbf{Q}^{-i}(t) = 0\}$ is Lipschitz continuous in F_j, in the the sense that

$$|P\{\mathbf{Q}^{-i}(t) = 0\} - P\{\tilde{\mathbf{Q}}^{-i}(t) = 0\}| \leq K \sup_{s\in[t_0,t]} |F_j(s) - \tilde{F}_j(s)|\,,$$

for some constant $K > 0$ and all $t \geq t_0$, where $\mathbf{Q}^{-i}$ and $\tilde{\mathbf{Q}}^{-i}$ correspond to F_i and $\tilde{F}_j$, respectively, and $F_j(s) = \tilde{F}_j(s)$ for $s < t_0$. The proof of this inequality essentially follows from Lemma 4. Evidently, the same continuity property holds when F_j is replaced by $F^{-i} = F_0 + F_j$. Similar observations clearly hold with i and j interchanged.

It follows that (35) is a retarded functional differential equation (in the sense of [4]) in the pair (F^{-i}, F^{-j}) over $t \geq t_0$, with initial conditions given at t_0, and Lipschitz continuous right hand side. Furthermore, (F^{-i}, F^{-j}) are continuous (given that they contain no point masses). It therefore follows by Theorem 2.2.3 in [4] that these equation has a unique solution in $[t_0, t_0 + \epsilon)$. But this solution must be symmetric, namely $F^{-i}(t) = F^{-j}(t)$, as otherwise we can obtain a different solution by interchanging the two. We have thus shown that $F_i(t) = F_j(t)$ for $t \in (t_0, t_0 + \epsilon)$, and the proof of case (ii) by contradiction is complete. □

Proposition 1 now follows as a direct consequence of Lemmas 15 and 12. □

C Proofs for Subsection 4.3

Proof of Lemma 9: (i) The Lipschitz continuity of P_0 in (14) clearly extends to the right-hand side of equation (12). Claim (i) now follows by by standard results, for example [4, Theorem 2.2.3].

(ii) Continuous dependence of $G(\cdot)$ on the initial conditions, uniformly in t, again follows from (14) by standard results, see [4, Theorem 2.2.2]. This clearly extends to $G'(t)$ by the Lipschitz continuity noted in that Lemma.

(iii) Consider two solutions G and F with $F(0) > G(0)$. Observe that $P_0(\mathbf{F_0}) = (1 - F(0))^M < (1 - G(0))^M = P_0(\mathbf{G_0})$, so that $F'(0) > G'(0)$, and by continuity this extends to $t \in [0, \epsilon]$ for some $\epsilon > 0$. Suppose now, by contradiction, that $F'(t) \leq G'(t)$ for some $t > \epsilon$, and let t_0 be the first time for which equality holds (which exists by continuity). But then $F'(t) > G'(t)$ for all $t \in [0, t_0)$, hence $F(t) > G(t)$ there. By Lemma 3(iii), this entails that $P_0(\mathbf{F_{t_0}}) < P_0(\mathbf{G_{t_0}})$, and the FDE implies that $F'(t_0) > G'(t_0)$, which contradicts the definition of t_0. □

The following notation will be used in the next proofs. Let

$$t_0 = \tau \text{ if } G'(\tau) = 0, \text{ and } \infty \text{ otherwise.}$$
$$t_1 = \tau \text{ if } G(\tau) = 1, \text{ and } \infty \text{ otherwise.}$$

(recall that τ is the first time at which either $G(t) = 1$ or $G'(t) = 0$.) Evidently, $\tau = \min\{t_0, t_1\}$.

Proof of Proposition 4: The proof proceeds in several steps.

1. Bounded final time: We first observe that τ is finite, and in fact uniformly bounded as a function of $G(0) \in [\gamma_{\min}, 1]$. Indeed, observe that the expected number of departures by

time t is given by (cf. (4)):

$$E(\mathbf{D}(t)) = \mu E(\mathbf{B}(t)) = \mu \int_0^t (1 - P_0(\mathbf{G_t}))dt\,.$$

But equation (12) implies that $P_0(\mathbf{G_t}) < \frac{\alpha}{\alpha+\beta}$ as long as $G'(t) > 0$, so that $E(\mathbf{D}(t)) \geq \mu\frac{\beta}{\alpha+\beta}t$. But since $\mathbf{D}(t)$ cannot exceed M, the number of potential arrivals, it follows that either $G'(t) = 0$ or $G(t) = 1$ must occur for some $t \leq \frac{M(\alpha+\beta)}{\mu\beta} \triangleq T_{\max}$. Therefore $\tau \leq T_{\max}$ for all $G(0)$.

2. Monotonicity of t_0: Suppose $\tau = t_0$ for some $G(0) = \gamma$, namely $G'(\tau) = 0$ and $G(\tau) \leq 1$. Observe that by Lemma 9(iii) both $G(t)$ and $G'(t)$ are strictly increasing in $G(0)$ for any fixed t. It follows directly that $\tau = t_0$ for all $G(0) < \gamma$, and further that t_0 is strictly increasing in $G(0)$ there.

3. Monotonicity of t_1: Suppose $\tau = t_1$ for some $G(0) = \gamma$, namely $G'(\tau) \geq 0$ and $G(\tau) = 1$. It similarly follows, by monotonicity of $G(t)$ and $G'(t)$ in $G(0)$, that $\tau = t_1$ for all $G(0) > \gamma$, and further that t_1 is strictly decreasing in $G(0)$ there.

4. Crossing to infinite t_0, t_1: Let $\gamma_0 = \sup\{G(0) \in [\gamma_{\min}, 1] : t_0 < \infty\}$, and $\gamma_1 = \inf\{G(0) \in [\gamma_{\min}, 1] : t_1 < \infty\}$. Observe that both t_0 and t_1 are indeed finite for some $G(0)$ in that range. Indeed, $t_0 = 0$ for $G(0) = \gamma_{\min}$ (since then $G'(0) = 0$), and $t_1 = 0$ for $G(0) = 1$.

5. $\gamma_0 = \gamma_1$: We can infer that $\gamma_0 = \gamma_1$ from the above-mentioned monotonicity properties of t_0 and t_1. Indeed: $t_0 < \infty$ for $G(0) < \gamma_0$; $t_1 < \infty$ for $G(0) > \gamma_0$; t_0 and t_1 cannot both be infinite for the same $G(0)$ since τ is finite; and finally note that if t_0 and t_1 are both finite then $t_0 = t_1$ by their definition, so that this can hold for a single value of $G(0)$ at most, due to the opposite monotonicity of t_0 and t_1 in $G(0)$.

6. Properties (i) and (ii): Let γ^* denote the common value of γ_0 and γ_1. Given the definitions of the latter, we obtain the following:
– For $G(0) < \gamma^*$, $t_1 = \infty$ and $\tau = t_0 < \infty$; that is, $G'(\tau) = 0$ and $G(\tau) < 1$.
– For $G(0) > \gamma^*$, $t_0 = \infty$ and $\tau = t_1 < \infty$; that is, $G'(\tau) > 0$ and $G(\tau) = 1$.
In view of the monotonicity properties of t_0 and t_1, it follows that properties (i) and (ii) of the Proposition are satisfied. □

Proof of Theorem 2: Consider γ^* from Proposition 4. Clearly both terminal conditions can hold together only for $G(0) = \gamma^*$. It remains to show that they indeed hold at γ^*.

Suppose $G(0) = \gamma^*$. Let t_0 and t_1 be defined as in the last proof. Recall that t_0 and t_1 cannot both be infinite (since τ is bounded), and $t_0 = t_1$ when both are finite. Therefore,

there are three mutually-exclusive possibilities at γ^*:
– Option 1: $t_0 = \infty$. That is, $\tau = t_1 < \infty$, $G(\tau) = 1$, $G'(\tau) > 0$.
– Option 2: $t_0 = t_1 < \infty$. That is, $G(\tau) = 1$, $G'(\tau) = 0$.
– Option 3: $t_1 = \infty$. That is, $\tau = t_0 < \infty$, $G(\tau) < 1$, $G'(\tau) = 0$.
Option 2 is just the required property in the theorem statement. Hence, the proof will be complete once we rule out Options 1 and 3.

In the following, we let $G(t, \gamma)$ and $\tau(\gamma)$ denote the solution and final time that correspond to $G(0) = \gamma$, and similarly for $t_0(\gamma)$, $t_1(\gamma)$. Also denote $\tau^* = \tau(\gamma^*)$. Recall that $G(t, \gamma)$ and $G'(t, \gamma)$ are continuous in t and γ (Lemma 9).

Suppose Option 1 holds at $G(0) = \gamma^*$. By Proposition 4(i), for $G(0) = \gamma < \gamma^*$ we have $G'(\tau(\gamma), \gamma) = 0$, $G(\tau(\gamma), \gamma) < 1$, and $\tau(\gamma) = t_0(\gamma)$ is increasing in γ there. Define $\hat{\tau} = \lim_{\gamma\uparrow\gamma^*} t_0(\gamma)$. Consider two cases.

a. $\hat{\tau} \leq \tau^*$. By continuity of $G'(t; \gamma)$ in γ and t, we obtain $G'(\hat{\tau}, \gamma^*) = \lim_{\gamma\uparrow\gamma^*} G'(t_0(\gamma), \gamma) = 0$. But since $\hat{\tau} \leq \tau^*$ this means that $t_0(\gamma^*) = \hat{\tau} < \infty$, contrary to Option 1.

b. $\hat{\tau} > \tau^*$. This means, in particular, that $\tau(\gamma) = t_0(\gamma) > \tilde{\tau}$ for some $\tilde{\tau} > \tau^*$ and all $\gamma < \gamma^*$ close enough to γ^*. However, we will use $G(\tau^*, \gamma^*) = 1$ and $G'(\tau^*, \gamma^*) > 0$ to show that $G(\tau^* + \epsilon, \gamma) > 1$ must hold for any $\epsilon > 0$ and $\gamma < \gamma^*$ close enough to γ^*, in obvious contradiction to $\tau(\gamma) > \tilde{\tau}$. Indeed, by continuity in γ, $\lim_{\gamma\uparrow\gamma^*} G'(\tau^*, \gamma) = G'(\tau^*, \gamma^*) \overset{\triangle}{=} a_0 > 0$. Therefore, there exists $\gamma_1 < \gamma^*$ so that $G'(\tau^*, \gamma_1) \geq \frac{1}{2}a_0$. Further, by continuity in t, there exists $\epsilon_1 > 0$ so that $G'(t, \gamma_1) \geq \frac{1}{4}a_0$ for $t \in [\tau^*, \tau^* + \epsilon_1]$. By monotonicity in γ, this inequality extends to all $\gamma \in [\gamma_1, \gamma^*)$. Therefore,

$$G(\gamma, \tau^* + \epsilon_1) \geq G(\gamma, \tau^*) + \frac{1}{4}a_0\epsilon_1 \quad \gamma \in [\gamma_1, \gamma^*)\,.$$

Now, since $\lim_{\gamma\uparrow\gamma^*} G(\gamma, \tau^*) = G(\gamma^*, \tau^*) = 1$, it follows that $G(\gamma, \tau^* + \epsilon_1) > 1$ for γ close enough to γ^*. But this holds for ϵ_1 arbitrarily small (and in particular for $\epsilon_1 < \hat{\tau}$), which obtains the required contradiction. This completes the argument that Option 1 is not possible.

Turning to Option 3, suppose that it holds as stated. That is $G(\tau^*, \gamma^*) < 1$, $G'(\tau^*, \gamma^*) = 0$. We will show that this contrasts with property (ii) of γ^* in Proposition 4, namely that $G(\tau(\gamma), \gamma) = 1$ and $G'(\tau(\gamma), \gamma) > 0$ for $G(0) = \gamma > \gamma^*$. Define $\hat{\tau} = \lim_{\gamma\downarrow\gamma^*} t_1(\gamma)$. Consider again two cases.

a. $\hat{\tau} \leq \tau^*$. Then $G(\tau^*, \gamma^*) \geq G(\hat{\tau}, \gamma^*) = \lim_{\gamma\downarrow\gamma^*} G(\tau(\gamma), \gamma) = 1$, which contradicts $G(\tau^*, \gamma^*) < 1$.

b. $\hat{\tau} > \tau^*$. Here, analogously to the argument in Option 1(b), we will show that $G'(\tau^* + \epsilon, \gamma) < 0$ must hold for $\epsilon > 0$ and $\gamma > \gamma^*$ close enough to γ^*, which contradicts $\hat{\tau} > \tau^*$. For that purpose we will require some properties of the second derivative G''. The argument proceeds through several claims.

(b1) For any γ, $G(t) < 1$ and $G'(t) = 0$ imply that $G''(t) < 0$.
Indeed, by (12) $G''(t) = -\frac{\mu}{N-1}\frac{d}{dt}P\{\mathbf{Q}^{-i}(t) = 0\}$. But $\frac{d}{dt}P\{\mathbf{Q}^{-i} = 0\} > 0$ follows similarly to Lemma 6.

(b2) $G''(t,\gamma)$ is locally Lipschitz continuous in t and γ, and uniformly so while $G(t,\gamma) \leq 1 - \epsilon < 1$.
To see this, consider the differential equation (8) jointly with the evolution equations for the queue size probabilities $p_t(k,m)$ in (18). Note that $P\{\mathbf{Q}^{-i}(t) = 0\} = \sum_{m=0}^{N-1} p_t(0,m)$. Together these equations can be considered as a set of ordinary differential equations, with the right-hand size smooth and uniformly bounded as long as $1 - G(t) \geq \epsilon > 0$. The conclusion now follows by expressing $G''(t)$ in terms of these variables.

(b3) $G''(t,\gamma)$ is strictly negative in some neighborhood of (τ^*, γ^*).
Indeed, recalling that $G(\tau^*, \gamma^*) < 1$ and $G'(\tau^*, \gamma^*) = 0$, we have by (b1) that $G''(\tau^*, \gamma^*) \overset{\triangle}{=} -b_0 < 0$. Further, by the Lipschitz continuity of G there is an $\epsilon_3 > 0$ so that $G(\tau^*, \gamma^*) < 1 - \epsilon_3$ for all $t \leq \tau^* + \epsilon_3$ and $\gamma \in [\gamma^*, \gamma^* + \epsilon_3]$. Now (b2) implies that G'' is uniformly Lipschitz in that region, so that $G''(t,\gamma) < -\frac{1}{2}b_0$ for (t,γ) as above, possibly with smaller $\epsilon_3 > 0$.

(b4) Estimating G': A two-term Taylor expansion for $G'(t)$ gives

$$G'(t,\gamma) = G'(\tau^*,\gamma) + G''(\zeta,\gamma)(t - \tau^*)$$

for some $\zeta = \zeta(t,\gamma) \in [\tau^*, t]$. Restricting to $t \in (\tau^*, \tau^* + \epsilon_3)$ and $\gamma \in (\gamma^*, \gamma^* + \epsilon_3)$ as in (b3), we get

$$G'(t,\gamma) \leq G'(\tau^*,\gamma) - \tfrac{1}{2}b_0(t - \tau^*)\,.$$

Now, since $\lim_{\gamma\downarrow\gamma^*} G'(\tau^*,\gamma) = G'(\tau^*,\gamma^*) = 0$, it follows that

$$\lim_{\gamma\downarrow\gamma^*} G'(t,\gamma) < 0$$

for $t \in (\tau^*, \tau^* + \epsilon_3)$. This provides the required contradiction to $\hat{\tau} > \tau^*$, so that Option 3 is ruled out.

We are therefore left with Option 2, which establishes the Theorem. □