האוניברסיטה העברית בירושלים

THE HEBREW UNIVERSITY OF JERUSALEM

STRATEGIC TIMING OF ARRIVALS TO A FINITE QUEUE MULTI-SERVER LOSS SYSTEM

By

MOSHE HAVIV and LIRON RAVNER

Discussion Paper # 675 (December 2014)

מרכז פדרמן לחקר הרציונליות

THE FEDERMANN CENTER FOR THE STUDY OF RATIONALITY

Feldman Building, Edmond J. Safra Campus, Givat-Ram, Jerusalem 91904, Israel
PHONE: [972]-2-6584135 FAX: [972]-2-6513681
E-MAIL: ratio@math.huji.ac.il
URL: http://www.ratio.huji.ac.il/

Strategic timing of arrivals to a finite queue multi-server loss system

Moshe Haviv and Liron Ravner*

Department of Statistics and the Federmann Center for the Study of Rationality
Hebrew University of Jerusalem

December 9, 2014

Abstract

We provide Game-theoretic analysis of the arrival process to a multi-server system with a limited queue buffer, which admits customers only during a finite time interval. A customer who arrives at a full system is blocked and does not receive service. Customers can choose their arrival times with the goal of minimizing their probability of being blocked. We characterize the unique symmetric Nash equilibrium arrival distribution and present a method for computing it. This distribution is comprised of an atom at time zero, an interval with no arrivals (a gap), and a continuous distribution until the closing time. We further present a fluid approximation for the equilibrium behaviour when the population is large, where the fluid solution also admits an atom at zero, no gap, and a uniform distribution throughout the arrival interval. In doing so, we provide an approximation model for the equilibrium behaviour that does not require a numerical solution for a set of differential equations, as is required in the discrete case. For the corresponding problem of social optimization we provide explicit analysis of some special cases and numerical analysis of the general model. An upper bound is established for the price of anarchy (PoA). The PoA is shown to be not monotone with respect to population size.

*moshe.haviv@gmail.com, lravner@gmail.com

1 Introduction

Game-theoretic analysis of queues often focuses on decisions made in steady-state conditions. Examples of such decisions are whether or not to join, choosing a queue, and bidding for priority in service admission. Hassin and Haviv provide a detailed survey of such models in [5]. The present work belongs to the narrower field of models where customers choose their time of arrival. This calls for transient analysis of the queueing process which is typically less tractable than the steady-state analysis. The motivation for the study of such models comes from services that are provided only for a certain time, such as the check-in procedure at an airport terminal or a medical clinic. Even if the service can be obtained at any time, the bulk of the demand may be during a certain time interval, for instance, a road leading to a business district that is utilized mainly during morning rush hour or a video-streaming website that is mostly visited in the evening hours. In some cases, such as the video-streaming website and medical clinic, the capacity of the queue may be limited, thus resulting in the blocking (or loss) of some of the arriving customers. This brings us to the goal of our work, which is to analyse a system where customers are interested in arriving at a time that maximizes the probability of being admitted into the system.

A queue with endogenous arrival times, denoted by $?/M/1$, was introduced by Glazer and Hassin in [3]. They modelled a non-cooperative game where a Poisson distributed number of customers time their arrival to a single-server queue with a limited admission interval, $[0, T]$. The objective of the customers is to minimize their own waiting times. The (symmetric) Nash equilibrium strategy was shown to be mixed. In particular, a continuous distribution that is uniform for some period before the opening, has a downward discontinuity at zero and a decreasing density function between zero and T.

In this work we analyse the $?/M/m/c$ model, which is a system of $m \geq 1$ identical exponential servers and a limited queue buffer of size $c \geq 0$. This is an extension of our previous research in [21], where the $?/M/1/0$ model was analysed. In our loss model there is no waiting time cost but rather a binary objective function: success in obtaining service, or failure to do so.

The analysis of strategic arrival times to a congested system has long been the subject of research in the transportation literature. Vickrey presented in [23] a game of timing arrivals to a bottleneck with the goal of minimizing waiting times. It was assumed that the population is fluid, that is to say, that

a single customer has infinitesimal influence on the system. This assumption is prevalent throughout the rich subsequent transportation literature on such models. The key difference between this model and the aforementioned queueing approach introduced in [3] is that in the latter the customers are discrete and thus every single arrival has a quantum impact on the system. A discrete population model that appears in the transportation literature is [18] by Otsubo and Rapoport. Unlike in the queueing literature, however, their model assumes that arrivals can occur only in discrete time intervals and that the service time is deterministic.

The research on queueing arrival-time games has evolved and examined various model assumptions over the years. We next review some of the main developments. Glazer and Hassin studied an arrival time game with batch service in [4]. Mazalov and Chuiko [17] assumed a single server with no queue buffer and customers who have a time sensitivity function that they wish to minimize instead of their own waiting costs. In [6] Hassin and Kleiner imposed a condition that customers cannot queue before the opening time and showed that in equilibrium there is an atom at time zero and a gap immediately there after. Customers may also incur tardiness costs on top of the waiting costs and such a model was examined by Jain, Juneja, and Shimkin [12]. Juneja and Shimkin [14] provided a rigorous characterisation of a general model for any distribution of the number of customers. They also provided a proof of the uniqueness of the equilibrium and proved that the equilibrium solution converges to that of the fluid model when the population size grows with the appropriate scaling of the other parameters. Haviv [8] considered a model combining tardiness costs, waiting costs and restrictions on opening and closing times, along with a fluid approximation. Honnappa and Jain [10] examined a fluid multi-server system where customers are allowed to pick a server as well as their arrival time. In [20] Ravner studied a model where the individual cost of the customers increases with the order of their admission. The equilibrium arrivals to a fluid queue with a processor-sharing regime was analysed by Juneja and Raheja in [13]. A more general analysis of the probabilistic properties of the transient queueing process prevailent in these models was carried out by Honnappa, Jain, and Ward in [11]. They characterised both the fluid and the diffusion limits of the process.

Controlling arrival times to queues with the goal of minimizing objectives such as waiting times, tardiness, or server idleness is a research topic that has been mostly orthogonal to the game-theoretic literature. The analysis usually requires different tools and is closer to the field of scheduling. Nevertheless,

the results of the equilibrium and the scheduling analysis complement each other and together provide deeper insights into the phenomenon of transient arrival processes to congested systems. The price of anarchy, which is defined as the ratio between the socially optimal utility and the social utility in equilibrium, is an interesting measure of the "damage" done by selfish customers. The minimization of waiting costs in queues by controlling the arrival times was studied by Pegden and Rosenshine in [19]. They showed that the objective function (which also includes a penalty for the idle time of the server) is convex for a small number of customers. It was conjectured that this is true for any number of customers, but as far as we know this remains a conjecture and has not been verified. Because of the difficulty of obtaining closed-form expressions of the transient moments of queues there is a need for an algorithmic approach. If the objective function is indeed convex then straightforward search algorithms can find the globally optimal schedule. Stein and Cote obtained an optimal equally spaced schedule in [22]; i.e., interarrival times are constant with an objective of finding the optimal constant. This approximation also appeared in the doctor appointment-scheduling context in the work of Hassin and Mendel [7], where they also accounted for no-shows. Altman et al. [1] provided general conditions for discretized local search methods to converge to a global optimum, namely, the multimodularity property. They also presented a queueing application for controlling arrivals to a $D/D/1$ queue with batch service. This method was later applied to the minimization of waiting times by Kaandorp and Koole in [15] and in the aforementioned work [6]. The latter analysed the optimal symmetric arrival strategy (identical arrival distribution for all customers) to a queue in a predefined interval. In [21] an explicit solution for the special case of our $?/M/1/0$ model was derived. The optimal schedule was shown to be a symmetric equally spaced grid on all of the arrival interval. Upper bounds for the price of anarchy in queueing arrival time games appeared in [14] and [20]. Explicit results were obtained for the fluid case in the former, and also in [8]. An optimization problem that is close to the one studied in this paper was presented in [16] by Lin and Ross. They sought an optimal dynamic "gatekeeper" policy for a partially observable loss server, where the optimizer observes the arrival times but not the state of the server.

The remainder of this paper is arranged as follows. In Section 2 we define the model and provide preliminary analysis of the queueing process and of the game. In Section 3 we summarize the single-server results of [21]. We proceed to characterise the unique symmetric Nash equilibrium of the general multi-

server model in Section 4. We show that the equilibrium arrival distribution has an atom at zero, an interval with no arrivals (a gap), and a continuous density on the remaining interval. This density stems from a set of functional differential equations defined by the transient probabilities of the underlying queueing process. We further provide a numerical method for computing the equilibrium arrival distribution. In Section 5 we present a fluid version of the game and derive the Nash equilibrium strategy, which is also socially optimal in this case. In Section 6 we present some numerical examples of both the discrete and the fluid approximation equilibrium solutions. Section 7 is devoted to the social optimization problem. We provide a general bound for the price of anarchy using an optimal dynamic control. We further discuss the difficulties of obtaining an optimal schedule for the general model. Using a simple example we illustrate that the individual customer loss probabilities are not convex with respect to the customer arrival times, and generally not symmetric under the optimal scheme. We present a numerical comparison of the social utility in equilibrium with that of an equally spaced symmetric schedule and of a global upper bound. Finally, in Section 8 we summarize our results and suggest future research avenues.

2 Model and preliminaries

In this section we first present the queueing model, followed by a formulation of the arrival game and the analysis of some of its basic properties. Throughout this work we make use of the following notations: the minimum and maximum of x and y are denoted by $x \wedge y$ and $x \vee y$, respectively. For $n \geq 1$ and $0 \leq p \leq 1$ we denote the binomial probability by $b(j; n, p) := \binom{n}{j} p^j (1-p)^{n-j}$, for $0 \leq j \leq n$, and the respective r.v. by $B_{n,p}$.

2.1 The queue

A system of $m > 0$ identical servers with exponential service rate μ and $c \geq 0$ waiting spots opens at time 0 and closes its gate at time T. Customers in the system at time T will complete their service. If there is at least one idle server upon a customer's arrival then his service commences immediately. Otherwise, if all servers are busy, the customer joins a *FCFS* queue if its length is less than c or leaves the system without obtaining service if there is no vacant waiting spot. If $n > 1$ customers arrive at exactly the same

time and there are only $k < n$ available spots, then the customers admitted into the system are determined by a uniform random draw. We denote the set of customers who wish to obtain service by $\mathcal{N} = \{1, \ldots, N\}$, where $N > 1$. The arrival times of the customers are independent random variables, denoted by $\mathcal{A} := \{A_1, \ldots, A_N\}$ with distributions (*cdf*) $\mathcal{F} := \{F_1, \ldots, F_N\}$, respectively. Note that the index of the customers is arbitrary and does not correspond to the actual order of arrival, which is determined by the order statistics $A_{[1]}, \ldots, A_{[N]}$. Thus, interarrival time $i \in \mathcal{N}$ is given by $\Delta(i) = A_{[i]} - A_{[i-1]}$, where $A_{[0]} = 0$. The arrival process is not a renewal process since the interarrival times are neither independent nor identically distributed.

Let $Q_{N,\mathcal{F}}(t)$ and $A_{N,\mathcal{F}}(t)$ denote the number of customers presently in the system and the number that have already arrived at time $t \in [0, T]$, respectively. The process $\{(Q_{N,\mathcal{F}}(t), A_{N,\mathcal{F}}(t)), t \in [0, T]\}$ is a continuous-time Markov chain with non-homogeneous in time transition rates that are determined by N and $\mathcal{F}$. If $F_i = F$, $\forall i \in \mathcal{N}$, then we simply write F instead of $\mathcal{F}$ in the process notation. Let $p_{ij}(t) := \mathrm{P}(Q_{N,\mathcal{F}}(t) = i, A_{N,\mathcal{F}}(t) = j)$ be the probability of state (i, j) at time $t \in [0, T]$. For the sake of a clearer presentation of the queuing dynamics, we omit N and $\mathcal{F}$ from the notation of $p_{ij}(t)$.

Remark The case of homogeneous F will be of particular interest in the equilibrium analysis. The generalization to different F_i is straightforward and can be found for an unlimited buffer model in [14]. Another straightforward generalization is allowing N to be a random variable. We briefly discuss this generalization for our model in Section 4.2. For further details on this case see, e.g, [21].

If all arrival times are indeed *iid* with F, then at any point t in which the density $f(t) = F'(t)$ is well defined the Kolmogorov backward equations

of the process are

$$\begin{aligned}
p_{0,j}^{'}(t) &= \mu p_{1,j}(t) - \lambda_j(t) p_{0,j}(t), \\
&\qquad 0 \le j \le N, \\
p_{i,j}^{'}(t) &= \lambda_{j-1}(t) p_{i-1,j-1}(t) + \mu_{i+1} p_{i+1,j}(t) - (\mu_i + \lambda_j(t)) p_{i,j}(t), \\
&\qquad 1 \le i \le j \wedge (m+c),\ j \le N, \\
p_{m+c,j}^{'}(t) &= \lambda_{j-1}(t)(p_{m+c,j-1}(t) + p_{m+c-1,j-1}(t)) - (\mu_m + \lambda_j(t)) p_{m+c,j}(t), \\
&\qquad m + c < j \le N,
\end{aligned} \tag{1}$$

where $\lambda_j(t) := (N-j)\frac{f(t)}{1-F(t)}$ and $\mu_i := (i \wedge m)\mu$. Note that $p_{ij}(t) = 0$ if $j < i$.

The loss probability is

$$\mathrm{P}(Q_{N,F}(t) = m + c) = \sum_{j=m+c}^{N} p_{m+c,j}(t) = \mathbb{E}\mathbb{1}_{\{A_{N,F}(t) - \sum_{j=1}^{N} \mathbb{1}_{\{A_j + W_j < t\}} = m+c\}}, \tag{2}$$

where W_j is the sojourn time of customer j (which can equal zero if the system is full upon his arrival). Note that if F is continuous at t then so is the loss probability, and since the service times are continuous, a jump can only occur when there is a jump in the arrival distribution.

We denote the expected number of customers yet to arrive at time t, given the state of the system, by $a_{N,F}(t;i) := \mathbb{E}(N - A(t) | Q(t) = i)$. The following lemma will be useful in the equilibrium analysis of Section 4.

Lemma 2.1 *For any t in the interior of the support of F,*

$$a_{N,F}(t;i) = N(1 - F(t)) \frac{\mathrm{P}(Q_{N-1,F}(t) = i)}{\mathrm{P}(Q_{N,F}(t) = i)}. \tag{3}$$

Proof

$$\begin{aligned}
a_{N,F}(t;i) &= \frac{\mathbb{E}(N - A_{N,F}(t))\mathbb{1}_{\{Q_{N,F}(t)=i\}}}{\mathrm{P}(Q_{N,F}(t)=i)} \\
&= N - \frac{\mathbb{E}\sum_{j=1}^{N}\mathbb{1}_{\{A_j \le t, Q_{N,F}(t)=i\}}}{\mathrm{P}(Q_{N,F}(t)=i)} \\
&= N\frac{\mathrm{P}(Q_{N,F}(t)=i) - \mathrm{P}(A_1 \le t, Q_{N,F}(t)=i)}{\mathrm{P}(Q_{N,F}(t)=i)} \\
&= N\frac{\mathrm{P}(A_1 > t, Q_{N,F}(t)=i)}{\mathrm{P}(Q_{N,F}(t)=i)} \\
&= N\frac{\mathrm{P}(Q_{N,F}(t)=i | A_1 > t)(1-F(t))}{\mathrm{P}(Q_{N,F}(t)=i)} \ .
\end{aligned}$$

If t is in the interior of the support of F then in the last line we condition on an event with positive probability. Given that $A_1 > t$, the queueing process is determined by the arrival and service times of the other $N-1$ customers, and since they are all independent, the proof is completed. ∎

2.2 The game

The parameters of the queueing game are $\langle N, m, c, \mu, T\rangle$, as defined above. Suppose the customers can choose their arrival time into the system. Specifically, customer j can choose the distribution F_j of A_j. A pure strategy assigns probability one to some time t; otherwise a mixed strategy is specified. We denote the support of strategy F_i by τ_{F_i}. A strategy profile $\mathcal{F} = \{F_i : i \in \{1, ..., N\}\}$ is the set of N strategies chosen by all customers, and $\mathcal{F}_{-i} := \mathcal{F} \setminus F_i$. A customer's utility is defined as the probability of obtaining service given the strategies of all customers. We denote the probability of customer j obtaining service when arriving at time t by $p_j^{\mathcal{F}_{-j}}(t)$. This probability equals $\mathrm{P}(Q_{N-1,\mathcal{F}_{-j}}(t) < m + c)$, as defined in the previous section.

Definition A strategy profile $\mathcal{F}$ is an equilibrium if there exists a constant C_i such that $p_i^{\mathcal{F}}(t) \le C_i$ for any customer i and any time t. Further, if $t \in \tau_{F_i}$ then $p_i^{\mathcal{F}}(t) = C_i$.

This indeed defines a Nash equilibrium since customers cannot choose any arrival time (or distribution) that will increase, in the strong sense, their

probability of obtaining service. From now on we will focus our analysis on symmetric equilibria where $F_i = F$, $\forall i \in \mathcal{N}$. We will show that such an equilibrium can always be constructed, although there may exist asymmetric equilibria as well. For example, in the game $\langle N = 2, m = 1, c = 0, \mu = 1, T = 1\rangle$, if customer 1 arrives at time 0 (respectively, 1) then customer 2's best response is arriving at time 1 (respectively, 0). This is because the probability of winning the random draw is lower than the probability of a service completion: $\frac{1}{2} < 1 - e^{-1}$. This yields two non-symetric and pure strategy equilibria, since the roles of the customer can be swapped. More generally, there exist $N!$ equilibria for every asymmetric equilibrium. Evidently, asymmetric equilibrium requires a distinction between the customers and a degree of cooperation. We argue that in large populations it is more natural to consider "anonymous" symmetric strategies. Note that a symmetric equilibrium may be pure. For instance, we will later provide conditions under which the arrival of all customers at time zero is an equilibrium. To avoid technicalities we limit the possible arrival distributions to those that satisfy the conditions for existence and uniqueness presented in [14]. These conditions rule out distributions that are hard to characterise and analyse such as distributions with a support that is defined by an infinite number of points and empty intervals. We next state some of the basic properties of a symmetric equilibrium. These arguments and results are along the lines of those presented in [6] and [8].

Consider an arbitrary customer, and let us assume that all other customers arrive at time zero with probability p. Denote the number of other customers arriving at time zero by $B_{N-1,p} \sim Bin(N-1, p)$. If there are m servers and a queue buffer of size c then the probability of obtaining service if he decides to arrive at time zero himself is $\mathbb{E}\left(\frac{m+c}{B_{N-1,p}+1} \wedge 1\right)$. The probability of obtaining service by arriving at time t, assuming all customers arrive with zero probability during the interval $(0, t]$, is

$$\mathrm{P}(B_{N-1,p} < m + c) + \mathrm{P}(B_{N-1,p} \geq m + c)\left(1 - e^{-m\mu t}\right).$$

This probability is decreasing with t, and as t approaches zero it approaches $\mathrm{P}(B_{N-1,p} < m + c)$. We can deduce that the probability of obtaining service

at time zero is higher than at the infinitesimal moment after zero:

$$\mathbb{E}\left(\frac{m+c}{B_{N-1,p}+1}\wedge 1\right) = \sum_{j=0}^{m+c-1} \mathrm{P}(B_{N-1,p}=j) + \sum_{j=m+c}^{N} \mathrm{P}(B_{N-1,p}=j)\frac{m+c}{j+1}$$
$$> \mathrm{P}(B_{N,p} < m+c).$$

In other words, it is better to join the random draw at zero than to arrive a moment later. Therefore, for any p there exists a unique t such that

$$\mathbb{E}\left(\frac{m+c}{B_{N-1,p}+1}\wedge 1\right) = 1 - \mathrm{P}(B_{N-1,p}\geq m)e^{-m\mu t}, \tag{4}$$

and the *rhs* is smaller than the *lhs* if t is replaced by $s \in (0,t)$.

Lemma 2.2 *If $N > m+c$ then a symmetric equilibrium arrival distribution F satisfies the following properties:*

1. *There is an atom at zero, $F(0) := p_e > 0$.*
2. *If $p_e < 1$ then there exists a time $0 < t_e < T$ such that $F(t_e) = F(0)$; there are no arrivals during the interval $(0, t_e)$. Furthermore, p_e and t_e satisfy equation (4).*
3. *There exists some positive density $f(t) := F'(t) > 0,\ \forall t \in [t_e, T]$.*

Proof 1. If $F(0) = 0$, contradicting property 1, then any customer can obtain service with probability one by arriving at time zero. This is only the case if the equilibrium probability of obtaining service is one, which is impossible because the service times are positive with probability one and $N > m+c$.

2. Let $F(0) := p_e < 1$ and let t_e be the respective solution to equation (4). We have shown that the probability of obtaining service by arriving at any time $s \in (0, t_e)$ is lower than at time zero, hence by Definition 2.2 this interval cannot be in the support of the equilibrium arrival distribution.

3. Finally, we argue that the distribution has no atoms and no holes in the interval $[t_e, T]$. Consider the loss probability, $\mathrm{P}(Q_{N-1,F}(t) = m+c)$ as defined in (2). An atom at t, namely $F(t-) < F(t)$, would imply an

upward jump in the loss probability, and so it is always better to arrive just before the atom. Had there been a hole in (s,t), $F(s) = F(t)$, where $t_e < s < t < T$, then the loss probability would have been lower at time t than at time s due to the possibility of a service completion, which would have contradicted the equilibrium assumption.

Lemma 2.3 *For any N, μ, m, and c, the time t that solves equation* (4) *is a decreasing function of p.*

Proof Rearranging equation (4), we get

$$e^{-m\mu t} = \frac{1 - \mathbb{E}(\frac{m+c}{B_{N,p}+1} \wedge 1)}{\mathrm{P}(B_{N,p} \geq m+c)} = \frac{\mathbb{E}(\frac{B_{N,p}+1-(m+c)}{B_{N,p}+1} \vee 0)}{\mathrm{P}(B_{N,p} \geq m+c)} = \mathbb{E}\left(\frac{B_{N,p}+1-(m+c)}{B_{N,p}+1} | B_{N,p} \geq m+c\right).$$

If $\{B_{N,p} | B_{N,p} \geq m+c\}$ is stochastically increasing with respect to p then the mean value of the increasing function $\frac{B_{N,p}+1-(m+c)}{B_{N,p}+1}$ is stochastically increasing as well. Thus, the larger p is, the larger $e^{-\mu t}$ is and hence the smaller t is. By taking the derivative of the function $\mathrm{P}(B_{N,p} \geq x | B_{N,p} \geq m+c)$ for any $x \geq m+c$, with respect to p, it can be shown that it is a strictly increasing function. ∎

At first glance the result of Lemma 2.3 might seem surprising, but careful consideration suggests it need not be. If in equilibrium a larger proportion of the customers arrive at time zero then the probability of obtaining service will be lower along all of the support. Hence, the fact that a shorter interval with no arrivals is required for the probability to increase to this lower probability is not unintuitive. On the other hand, it is not an obvious result. It will play an important role in establishing the equilibrium properties of the model.

Remark In [21] a similar result was proved in the case where $m = 1$ and $c = 0$, for any distribution of N. This generalization is straightforward for $m \geq 1$ and $c \geq 0$ as well.

3 Single-server and no queue buffer (a brief review)

In this section we briefly summarize the equilibrium analysis that was carried out in [21] for the case of $m = 1$ and $c = 0$ (using our notation). No proofs are supplied here. The unique symmetric equilibrium strategy satisfies:

- There exists a positive probability of arriving at time zero, $p_e := F(0) > 0$.
- There are no arrivals in the interval $(0, t_e)$, where

$$t_e = -\frac{1}{\mu} \log \left(\frac{(N-1)p_e - (1-(1-p_e)^{N-1})}{(N-1)p_e(1-(1-p_e)^{N-2})} \right).$$

[1]

- If for $p_e = 1$ the solution to (3) yields $t_e \geq T$ then $p_e = 1$; i.e., all customers arrive at time zero.
- Otherwise, $p_e < 1$ and the continuous arrival distribution is defined using the functional differential equation:

$$f(t) = \frac{\mu}{N} \cdot \frac{\mathrm{P}(Q_{N-1,F}(t) = 1)}{\mathrm{P}(Q_{N-2,F}(t) = 0)}, \ t \in [t_e, T]. \tag{5}$$

- The probability of obtaining service for any customer is $\frac{1-(1-p_e)^N}{Np_e}$, and the expected number of served customers is $\frac{1-(1-p_e)^N}{p_e}$.
- For $N = 2$, if $p_e < 1$ the equilibrium arrival distribution is uniform with $p_e = \frac{2}{2+T\mu-\log 2}$, $t_e = \frac{\log 2}{\mu}$, and $f(t) = \frac{\mu p_e}{2}, \ t \in [t_e, T]$.
- The distribution along $[t_e, T]$ is uniform when the number of participating customers follows a Poisson distribution.
- For the general population size a numerical method can be applied to solve equation (5) using a special case of the queueing dynamics presented here in (1).
- Numerical analysis suggests that the arrival distribution is "close" to uniform for large values of N.

[1] This is the solution to the equilibrium condition (4), presented in the previous section.

4 Multiple servers and a limited buffer

Let us now return to the general model with $m \geq 1$ identical servers, and a limited queue buffer of size $c \geq 0$. We assume that not all customers will be served if they all arrive at once, i.e. $N > m + c$. If this were not the case, then any arrival profile would be trivially a Nash equilibrium. We first characterise a symmetric equilibrium arrival distribution, followed by a proof of its uniqueness and a numerical procedure to compute it.

Theorem 4.1 *Let*

$$t(p) := -\frac{1}{m\mu} \log \left(\frac{1 - \mathbb{E}\left(\frac{m+c}{B_{N-1,p}+1} \wedge 1 \right)}{\mathrm{P}(B_p \geq m + c)} \right). \tag{6}$$

If $t(1) > T$ *then* $p_e = 1$. *Otherwise,* p_e *and* $t_e = t(p_e)$ *solve equations* (6) *and*

$$\int_{t_e}^{T} f(t)\ dt = 1 - p_e, \tag{7}$$

where f *satisfies the functional differential equation,*

$$f(t) = \frac{m\mu}{N-1} \cdot \frac{\mathrm{P}(Q_{N-1,F}(t) = m + c)}{\mathrm{P}(Q_{N-2,F}(t) = m + c - 1)} \mathbb{1}_{\{t \in [t_e, T]\}}. \tag{8}$$

Then

$$F(t) = p_e + \int_{t_e}^{t} f(s) ds \mathbb{1}_{\{t \geq t(p_e)\}},\ t \in [0, T],$$

is a Nash equilibrium.

Proof The relation between p_e and t_e is a direct result of Lemma 2.2. Without loss of generality we single out customer N, and analyse the queueing process given that the $N - 1$ other customers arrive according to F. His probability of obtaining service is equal at times zero and t_e. For F to be an equilibrium we further require that this probability remains constant on the interval $[t_e, T]$, i.e.,

$$\mathrm{P}(Q_{N-1,F}(t) = m + c) = C,\ t \in [t_e, T],$$

for some $0 < C < 1$ (given p_e we can compute C, but this is not required for the remainder of the proof). Using the notation of Section 2.1 for $N -$

1 customers arriving according to F, we can state the above equilibrium condition as $\sum_{j=m+c}^{N-1} p'_{m+c,j}(t) = 0$. According to the process dynamics given in (1) we get

$$\sum_{j=m+c}^{N-1} [\lambda_{j-1}(t)(p_{m+c,j-1}(t) + p_{m+c-1,j-1}(t)) - (\mu_m + \lambda_j(t))p_{m+c,j}(t)] = 0.$$

Recall that $\lambda_j = (N-1-j)\frac{f(t)}{1-F(t)},\ 0 \le j \le N-1$ and $\mu_i = (i \wedge m)\mu,\ 1 \le i \le m+c$. After some algebra it can be verified that this condition is equivalent to

$$\frac{f(t)}{1-F(t)} \sum_{j=m+c-1}^{N-2} (N-1-j)p_{m+c-1,j}(t) = m\mu \sum_{j=m+c}^{N-1} p_{m+c,j}(t).$$

This in fact translates the equilibrium condition to a balancing equation. The sum on the *rhs* is exactly the loss probability $\mathrm{P}(Q_{N-1,F} = m+c)$. If we denote the sum on the *lhs* by $l(t)$ then using (3) from Lemma 2.1 we have that for $t < T$,

$$\begin{aligned} l(t) &= a_{N-1,F}(t; m+c-1)\,\mathrm{P}(Q_{N-1,F}(t) = m+c-1) \\ &= (N-1)\,\mathrm{P}(Q_{N-2,F}(t) = m+c-1)(1-F(t)). \end{aligned}$$

This leads to $f(t)$ as defined in (8). Note that the numerator and denominator are both continuous, strictly positive, and well defined for any $t \le T$, so even though the conditional probability $\mathrm{P}(Q_{N-1,F}(t) = m+c-1 | A_1 > t)$ is not defined for $t = T$, we can nevertheless define

$$f(T) := \lim_{t \to T} \frac{m\mu}{N-1} \cdot \frac{\mathrm{P}(Q_{N-1,F}(t) = m+c)}{\mathrm{P}(Q_{N-2,F}(t) = m+c-1)}.$$

∎

Corollary 4.2 *The equilibrium probability of obtaining service is*

$$q_e := \mathbb{E}\left(\frac{m+c}{B_{p_e,N-1}+1} \wedge 1\right), \tag{9}$$

and Nq_e is the social utility in equilibrium.

Lemma 4.3 *Any function $G(t)$ that solves the equilibrium conditions* (6) *and* (8) *in Theorem 4.1, is increasing with respect to the initial condition $p = F(0)$, for any $t \in [0, T]$.*

Proof Let $0 < p < q \leq 1$, and denote by t_p and t_q the respective solutions to (6). Further denote by G_p and G_q the respective solutions of (8). From Lemma 2.3 we know that $t_p > t_q$, which implies that

$$G_q(t_p) > G_q(t_q) = q > p = G_p(t_p).$$

Thus, the claim is clearly true for $0 \leq t \leq t_p$. Recall that in equilibrium the loss probability is equal throughout the support, in particular at time zero. If more customers arrive at time zero, then the loss probability is higher for the process with q at any time $[t_q, T]$, than for the process with p. Now let us suppose there exists some $s > t_p$ such that $G_p(s) = G_q(s)$ and $G_p(t) < G_q(t),\ \forall t < s$; then $g_p(s) > g_q(s)$ and $\frac{g_p(s)}{1-G_p(s)} > \frac{g_q(s)}{1-G_q(s)}$. That is, the arrival rate is higher for $p > q$ at s while the service rate remains the same in both cases. Hence, the loss probability for at least one of them, q or p, cannot remain constant, which contradicts the equilibrium condition leading to (8). ∎

Lemma 4.4 *The symmetric equilibrium arrival distribution F in Theorem 4.1 is unique.*

Proof The uniqueness is a result of Lemmata 2.3 and 4.3, which state that t_e is monotone decreasing and $F(t)$ is monotone increasing with p_e, respectively. Therefore, the final equilibrium condition (7), $\int_{t_e}^{T} f(t)\ dt = 1 - p_e$, has a unique solution p_e, and so do t_e and $\{f(t),\ t \in [t_e, T]\}$. ∎

4.1 A numerical procedure

We now present a numerical procedure to compute the equilibrium arrival distribution. It is essentially a standard discrete approximation of a set of differential equations, which is similar to those used in [3] and many of the subsequent works discussed in the Introduction.

We first state the initial probabilities of the process $(Q_{N-1,F}(t_e), A_{N-1,F}(t_e))$, given the value of p_e. Let $S_j(t_e)$ be the random number of service completions during the interval $[0, t_e]$, given that j customers arrived at time zero. The probability of state (i, j) at time t_e is

$$p_{i,j}(t_e) = b(j; N-1, p_e)\, \mathrm{P}(S_j(t_e) = j - i),\ 0 \leq i \leq j \leq N-1. \qquad (10)$$

Observe that $\mathrm{P}(S_j(t_e) = j-i) = b(i; j, e^{-\mu t_e})$, for any $0 \le i \le j \le m$, because all arrivals commence service immediately at time zero and $S_j(t_e)$ counts how many complete it before time t_e. If $j > m$ then only m commence service straightaway and up to c additional customers form a queue. Let $V_k(j)$ denote the time between service completions $k-1$ and k for $k = 1, \ldots, j \wedge (m+c)$. For notational convenience we also define $V_0(j) := 0$ and $V_{j+1}(j) := \infty$. Since all service times are iid exponential random variables with rate μ,

$$V_k(j) \sim \begin{cases} \exp(m\mu), & 1 \le k \le j \wedge m \\ \exp((j \wedge (m+c) - k + 1)\mu), & m < k \le j \end{cases}.$$

Hence, the *pdf* of the number of service completions up to time t_e is

$$\mathrm{P}(S_j(t_e) = s) = \begin{cases} b(s; j, 1 - e^{-\mu t_e}), & 0 \le s \le j \le m \\ \mathrm{P}\left(\max\{r : \sum_{k=0}^{r} V_k \le t_e\} = s\right), & 0 \le s \le j \wedge (m+c), \\ & m \le j \le N \end{cases}. \tag{11}$$

In order to compute the second probability we consider the sum of all service times had j customers arrived; $\sum_{k=1}^{n_j} V_k + \sum_{k=n_j+1}^{m_j} V_k$, where $n_j := j \wedge m$ and $m_j := j \wedge (m+c)$. The first sum is distributed $Erlang(n_j, m\mu)$ and the second (which may be empty) is a convolution of exponential r.v.s with different rates for which we can explicitly compute the density (see for example p.40 in [2]). To compute the probability of the event

$$\left\{\max\{r : \sum_{k=0}^{r} V_k \le t_e\} = s\right\} = \left\{\sum_{k=0}^{s} V_k \le t_e, \sum_{k=0}^{s+1} V_k > t_e\right\},$$

we can use the law of total probability by conditioning on the value of the sum until n_j.

After computing the probabilities for all states of the processes for $N-1$ and $N-2$, we are ready to compute $f(t_e)$ according to (8). We proceed to approximate the two probabilities and density discretely in small increments of $\Delta > 0$, throughout the interval $(t_e, T]$. This is done incrementally using (1),

$$p_{i,j}(t + \Delta) = p_{i,j}(t) + \Delta p'_{i,j}(t).$$

We summarize how the above analysis can be used in a simple procedure in order to compute the equilibrium arrival distribution. The input for the algorithm is the game parameters, the discretization parameter $\Delta > 0$, and an accuracy parameter $\epsilon > 0$.

Algorithm 1

(1) Set the search range $\underline{p} = 0$ and $\overline{p} = 1$.

(2) Set $p_e = \overline{p}$ and compute t_e using (6). If $t_e > T$, stop and return $F(t) = F(0) = 1, \forall t \in [0, T]$.

(3) Set $p_e = \frac{1}{2}(\underline{p} + \overline{p})$.

(4) Compute t_e, $\mathrm{P}(Q_{N-1,F}(t_e) = m + c)$, $\mathrm{P}(Q_{N-2,F}(t_e) = m + c - 1)$, and $f(t_e)$ using (6), (10), (11), and (8).

(5) Compute $\mathrm{P}(Q_{N-1,F}(t) = m + c)$, $\mathrm{P}(Q_{N-2,F}(t) = m + c - 1)$, $f(t)$. and $F(t)$ in increments of $\Delta > 0$, until either one of the following two conditions is met:

   (a) $|F(t) - 1| < \epsilon$

   (b) $|t - T| < \epsilon$

(6) If both conditions are met then stop and return F. If only condition (a) is met then set $\overline{p} = p_e$ and return to (3). If only condition (b) is met then set $\underline{p} = p_e$ and return to (3).

Lemma 4.5 *Algorithm 1 converges to the unique equilibrium in a finite number of iterations. A single iteration of the algorithm requires at most* $(2(m + c)N + 1)\frac{T}{\Delta}$ *computations.*

Proof The convergence is guaranteed by the monotone behaviour shown in Lemma 4.3. The upper bound for the number of computations is a crude one and relies on the following facts: the discrete number of increments of size Δ in the interval $[t_e, T]$ is smaller than $\frac{T}{\Delta}$, the number of possible states is smaller than $N(m+c)$, and in every increment the probabilities are computed for two processes, one with $N - 1$ and one with $N - 2$, along with a single computation of the density. ∎

4.2 A random number of arrivals

Suppose now that customers share a common belief that the number of the customers is a random variable, N.[2] The analysis of Section 2.2 carries over with two changes:

- the joining rate $\lambda_j(t)$ now equals $\mathbb{E}(N-j|N \geq j)\frac{f(t)}{1-F(t)}$.
- the state variable j is bounded by the largest possible value for N (which might be infinity).

The case where N follows a Poisson distribution comes with a considerable simplification. Suppose N follows a Poisson distribution with parameter λ. Then, $\lambda_j(t)$ can be replaced with $\lambda f(t)$, which is not a function of j, thus making this state variable redundant. Denote by $p_i(t)$ the probability that i customers are present at time t, $0 \leq i \leq m+c$, $0 \leq t \leq T$. We then get the set of differential equations:

$$p_0'(t) = -\lambda f(t) p_0(t) - \mu p_1(t),\ t_e \leq t \leq T,$$
$$p_i'(t) = \lambda f(t) p_{i-1}(t) - (\lambda f(t) + \mu_i) p_i(t) + \mu_{i+1} p_{i+1}(t),\ t_e \leq t \leq T, i \geq 1.$$

The equilibrium arrival density is now given by the solution to

$$f(t) = \frac{m\mu}{\lambda} \cdot \frac{p_{m+c}(t)}{p_{m+c-1}(t)},\ t_e \leq t \leq T.$$

The solution can be obtained numerically as was prescribed in the case where N is deterministic. Note that since the process is one-dimensional the procedure requires fewer computational steps.

5 Fluid model

Let $\Lambda > 0$ be a volume of fluid customers arriving to a system in which every customer can be served in infinitesimal time, but a unit of customer volume requires μ service time from a single server. As before, the number of servers is m, the queue buffer can hold a volume of c, and the system only admits customers during the interval $[0, T]$. To avoid trivialities we assume $\Lambda > m\mu T + c$; i.e., not all users can be served even when the servers

[2] For a discussion how such a common belief can emerge see [9].

are fully utilized throughout the interval $[0, T]$. If F is a symmetric arrival distribution then the arrival process is deterministic, $\{\Lambda F(t) : t \in [0, T]\}$. At any time t such that $F(t)$ is continuous the arrival rate is $\Lambda f(t)$, where f is the density. Note that while the system's behaviour is deterministic, the identity of the customers admitted into service may not be, because a random draw is carried out at any time where the flow of arrivals surpasses the flow of departures.

Lemma 5.1 *If $\Lambda > m\mu T + c$, then the symmetric equilibrium arrival distribution F satisfies the following properties:*

1. *The probability of obtaining service for any customer is strictly between zero and one.*
2. *There exist no s and t, where $0 \leq s < t \leq T$ such that $F(t) = F(s)$. In particular, $f(t) > 0$ and $0 < F(t) < 1,\ \forall\ t \in (0, T)$.*
3. *The actual service rate is $m\mu$ at any time $t \in [0, T]$ and the buffer is always full.*
4. *There are no jumps, i.e., $F(t-) = F(t),\ \forall\ t \in (0, T)$.*
5. *The probability of obtaining service when arriving at time t is $\frac{m\mu}{\Lambda f(t)}$.*

Proof 1. If the server is fully utilized then $m\mu T$ customers are served in the interval $[0, T]$. The customers left in the buffer at time T will be served as well. Since the buffer size is c, a volume of at most $m\mu T + c$ can be served, and hence if $\Lambda > m\mu T + c$ then no F can guarantee that all the customers will be served, namely, $q_e < 1$. Trivially, no policy can guarantee that all users be blocked with probability one.

2. If $F(t) = F(s)$, where $0 \leq s < t \leq T$, then there is a period with no arrivals and a positive flow of departures, which means that the probability of obtaining service at time t is higher than at s, which contradicts the equilibrium condition.

3. If at time t the service rate is not $m\mu$ or the buffer is not full, then all arrivals at time $t-$ obtain service with probability one, contradicting property 1.

4. According to property 3, the buffer is full at any time t, which means that if there is a positive atom at time t then a volume of customers the size of the atom will not obtain service, once again contradicting property 1.

5. Combining properties 1–4, we have that the arrival distribution is continuous with no atoms, and the system is always full. Therefore, the proportion of customers obtaining service at any time equals the ratio of the service rate $m\mu$ and the arrival stream $\Lambda f(t)$.

Theorem 5.2 *The equilibrium arrival distribution is uniform on the interval* $[0, T]$. *If* $c > 0$ *then there is an atom at time zero. Otherwise, there are no atoms. The equilibrium cdf is*

$$F(t) = \frac{m\mu t + c}{m\mu T + c},\ t \in [0, T] \tag{12}$$

and the probability of obtaining service is $q_e = \frac{m\mu T + c}{\Lambda}$.

Proof The uniform density is an immediate result of property 5 in Lemma 5.1, together with the equilibrium condition that the probability of obtaining service is constant on all of the support: $q_e = \frac{m\mu}{\Lambda f(t)},\ \forall\ t \in [0, T]$. The probability of obtaining service at time zero is the proportion of arrivals that will be selected to join the buffer, $\frac{c}{\Lambda F(0)}$. Both of the above arguments give F as defined in (12). Clearly, if $c = 0$ then $F(0) = 0$ and $f(t) = \frac{1}{T},\ \forall\ t \in [0, T]$. ∎

Remark The equilibrium arrival distribution is socially optimal in the fluid case. The server is being fully utilized and therefore the expected number of customers served cannot be increased by using any other arrival schedule.

6 Numerical analysis

In Figures 1 and 2 we computed the density for different values of N, with $\mu_N = \frac{N}{10}$ and $m = 5$, for the case with a positive buffer ($c_N = \frac{N}{5}$) and the case with no buffer ($c = 0$), respectively. We can see that the distribution becomes more and more uniform on all of the arrival interval when the population size increases.

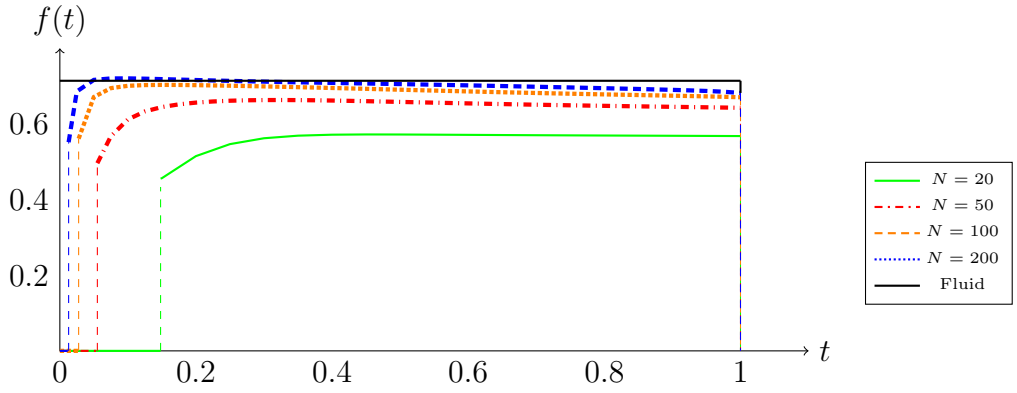

Figure 1: Equilibrium arrival density for $N \in \{20, 50, 100, 200\}$ with $\mu = \frac{N}{10}$, $c = \frac{N}{5}$, $m = 5$, and $T = 1$.

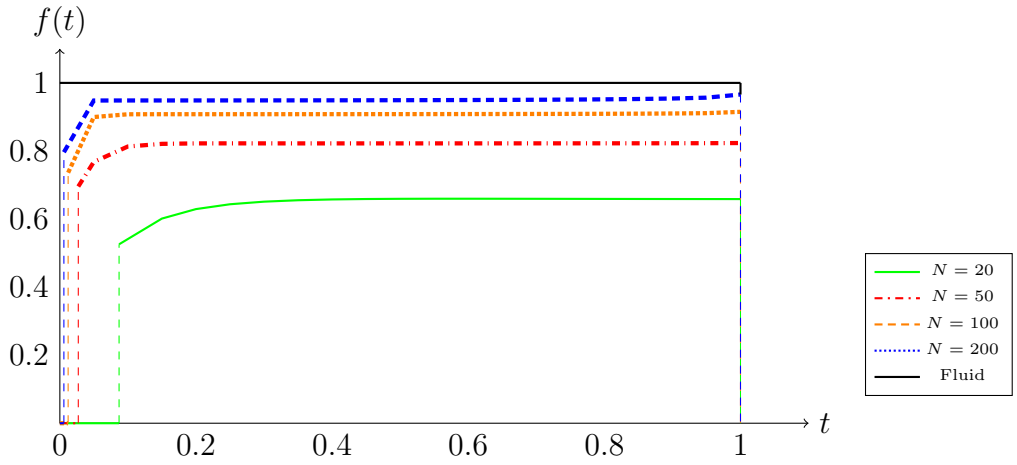

Figure 2: Equilibrium arrival density for $N \in \{20, 50, 100, 200\}$ with $\mu = \frac{N}{10}$, $c = 0$, $m = 5$, and $T = 1$.

The corresponding probabilities of arriving at zero and of obtaining service for both of the above examples are shown in Tables 1 and 2, respectively. Recall that $q_e = \mathbb{E}\left(\frac{m+c}{B_{N-1,p_e}+1} \wedge 1\right)$, where $B_{N-1,p_e} \sim Bin(N-1, p_e)$.

	$N = 20$	$N = 50$	$N = 100$	$N = 200$	Fluid
p_e	0.523	0.383	0.336	0.309	0.286
q_e	0.828	0.774	0.743	0.72	0.7

Table 1: Equilibrium probabilities of arriving at time zero and of obtaining service, for $N \in \{20, 50, 100, 200\}$ with $\mu_N = \frac{N}{10}$, $c_N = \frac{N}{5}$, $m = 5$, and $T = 1$.

	$N = 20$	$N = 50$	$N = 100$	$N = 200$	Fluid
p_e	0.402	0.195	0.104	0.054	0
q_e	0.613	0.509	0.481	0.463	0.5

Table 2: Equilibrium probabilities of arriving at time zero and of obtaining service, for $N \in \{20, 50, 100, 200\}$ with $\mu_N = \frac{N}{10}$, $c_N = 0$, $m = 5$, and $T = 1$.

In Figure 3 we see the equilibrium densities for values similar to those in the previous two examples, but with a scaling of the number of servers as a function of N while keeping a constant service rate, $\mu = 5$. The change in the arrival density is much smaller in this case, and it seems to be decreasing with the population size, rather than increasing as was the case in the previous examples. The changes in the equilibrium probabilities are also very small, as seen in Table 3. In any case, the distribution becomes more flat and uniform on all of the interval when the population size increases.

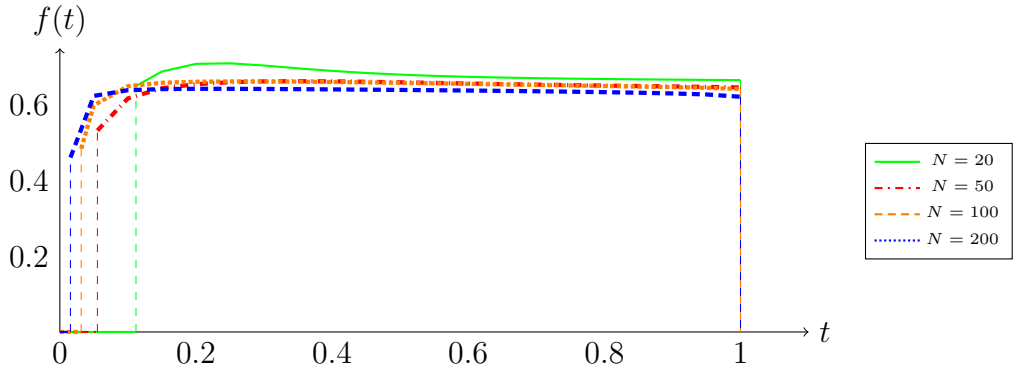

Figure 3: Equilibrium arrival density for $N \in \{20, 50, 100, 200\}$ with $\mu = 5$, $c = \frac{N}{5}$, $m = \frac{N}{10}$, and $T = 1$.

	$N = 20$	$N = 50$	$N = 100$	$N = 200$
p_e	0.398	0.382	0.371	0.379
q_e	0.728	0.775	0.805	0.791

Table 3: Equilibrium probabilities of arriving at time zero and of obtaining service, for $N \in \{20, 50, 100, 200\}$ with $\mu = 5$, $c = \frac{N}{5}$, $m = \frac{N}{10}$, and $T = 1$.

7 Social optimization

The utility of a customer was defined to be his probability of obtaining service given the arrival profile of all customers (including himself). Hence, the overall social utility is the expected number of customers admitted into service. The price of anarchy (PoA) is defined as the ratio between the optimal overall social utility and that of the equilibrium with the lowest social utility. As we have done throughout this paper, we consider only the symmetric equilibrium, which was shown to be unique. There may be additional equilibria with a lower expected number of customers served, which would lead to a higher PoA. The socially optimal policy may also have several definitions, which depend on the level of control that the central planner has over the customers. Three possible levels are: (1) the planner observes the queueing

process and can dynamically (i.e., online) assign arrivals, (2) the planner decides on a pre-determined (i.e., offline) arrival schedule, telling each customer exactly when to arrive, and (3) the planner can specify a single (symmetric) strategy for all customers. All three levels may be reasonable in different applications and have been studied in the scheduling literature, along with other control options not mentioned here. In the remainder of this section we use the first level of control to obtain an upper bound for all possible social utilities, provide more detailed analysis of the second level and briefly discuss the third.

For the arrival time game with parameters $\langle N, m, c, \mu, T\rangle$, we denote the equilibrium utility by $U^e(N)$ and the socially optimal utility by $U^*(N)$. The price of anarchy is then defined to be

$$PoA(N) := \frac{U^*(N)}{U^e(N)}. \tag{13}$$

Regardless of the policy adopted by the central planner, at least $m + c$ customers surely obtain service. Therefore, we seek a policy that maximizes the expected number of additional customers who are admitted into service. From the analysis of Section 4, we know that the equilibrium social utility is $U^e(N) = N\mathbb{E}(\frac{m+c}{B_{N-1,p_e}+1} \wedge 1)$, where p_e is the equilibrium probability of arriving at time zero. If the central planner can dynamically assign arrivals upon observing a service completion, then full system utilization is achieved for any realization of the service times. Therefore, the socially optimal utility is upper bounded by

$$\overline{U}^*(N) := m + c + \mathbb{E}\left[Y_{(m\mu T)} \wedge (N - m - c)\right],$$

where $Y_{(m\mu T)} \sim Poisson(m\mu T)$. Observe that if N is large then $\overline{U}^*(N) \approx m + c + m\mu T$. Also note that $\overline{U}^*(N)$ is not the accurate expected number of customers served by the optimal dynamic policy, because if more than $N - m$ customers are served then the last service completions will be at a slower rate than $m\mu$. We nevertheless find it more convenient to use this simpler bound. Clearly, any predetermined schedule, be it symmetric or not, cannot achieve a higher social utility than the optimal dynamic control. This leads to the following lemma.

Lemma 7.1 *For an arrival game with parameters $\langle N, m, c, \mu, T\rangle$,*

$$PoA(N) \leq \frac{\overline{U}^*(N)}{N\mathbb{E}(\frac{m+c}{B_{N-1,p_e}+1} \wedge 1)}. \tag{14}$$

Using an optimal dynamic policy to compute an upper bound on the price of anarchy also appeared in [14] and [20], where the goal was to minimize the waiting, tardiness, and order costs incurred by the customers. When the central planner is limited to specifying the customer's offline arrival strategies, the scheduling problem is typically hard and requires heuristic or approximation procedures. The objective is to specify a vector of interarrival times $\underline{t} = (t_1, \ldots, t_N)$ such that $\sum_{i=1}^{N} t_i \leq T$, which minimizes the expected number of blocked customers. It is clear that any optimal schedule assigns $m+c$ arrivals at time zero, as any schedule that does not do so can be trivially improved by moving to zero the first arrival after zero (who is admitted with probability one anyway). We are then left with choosing the inter-arrival times of the remaining $M := N - m - c$ customers, denoted by $t_1, \ldots, t_M$. It is also clear that the last arrival will always be scheduled at T, as again a trivial improvement can be obtained otherwise. The space of possible schedules can therefore be defined as $\mathcal{T} := \left\{ \underline{t} \in \mathbb{R}_+^M : \sum_{i=1}^{M} t_i = T \right\}$. For a given schedule $\underline{t} \in \mathcal{T}$, we denote the probability of the i'th arrival to be blocked by $\ell_i(\underline{t})$. The optimization problem is then

$$\min_{\underline{t} \in \mathcal{T}} \sum_{i=1}^{M} \ell_i(\underline{t}).$$

Let $s_i := \sum_{j=1}^{i} t_j$ and denote the number of customers in the system at time s when schedule $\underline{t}$ is being used by $Q_{\underline{t}}(s)$. The probability that arrival i is blocked is then

$$\ell_i(\underline{t}) = e^{-m\mu t_i} \, \mathrm{P}\left(Q_{\underline{t}}(s_{i-1}-) \in \{m+c-1, m+c\} \right). \tag{15}$$

For the special case of a single server and no queue buffer loss model ($m = 1, c = 0$) there are only two possible server states, busy and idle. We then have $\mathrm{P}\left(Q_{\underline{t}}(s) \in \{0,1\}\right) = 1$ for any $s \in [0,T]$. Thus, the optimization problem is reduced to a minimization of a sum of identical convex functions of the form $e^{-\mu t_i}$, constrained by $\sum_{i=1}^{M} t_i = T$. This was used in [21] to find an explicit solution for the social optimization problem. The optimal schedule is for the customers to arrive in equally spaced time intervals of length $\frac{T}{N-1}$, which results in a social utility of

$$U_N^* = N - (N-1)e^{-\frac{\mu T}{N-1}}.$$

It is interesting to point out that the globally optimal schedule is "equally spaced", as defined by Stein and Cote in [22] and further analysed by Hassin and Mendel in [7]. For the general case of $m \geq 1$ and $c \geq 0$, the objective function is more complex. There may be several service completions during any interval, $[t_{i-1}, t_i]$, implying that the probability of customer i being blocked depends on the state probabilities at several of the previous arrival instances. In order to compute the blocking probabilities we need to use the convolution of exponential distributions that was detailed in (11) for the initial state probabilities in the equilibrium computation. Note that this computation is required for every interval $t_i,\ i = 1, \ldots, M$. We next define a useful generalization of the equally spaced schedule.

Definition For $1 \leq k \leq m + c$, an $s(k)$-schedule is defined as follows. If $M \leq k$, set all M customers to time T. Otherwise, assign k customers to time T and assign the remaining customer arrivals on an equally spaced grid of the interval $[0, T]$ (with interval sizes of $\frac{T}{M-k+1}$).

We denote the expected utility achieved by the general equally spaced schedule by $U_N^{s(k)}$ and use it as a lower bound for the optimal utility and for the PoA. We will later present numerical analysis that suggests that this schedule is indeed closer to the socially optimal schedule than to the equilibrium utility (for any chosen $k > 0$). We will first present explicit examples where this schedule is optimal for different values of k.

Example: $M = 3$

Suppose that $m \geq 1$, $c \geq 1$, and that all but three customers can be admitted into the system together ($M = 3$). In Theorems 7.2 and 7.3 we present the explicit optimal solutions for two special cases that highlight the following interesting properties of the objective function:

1. The individually loss probabilities are generally not convex.

2. The expected number of lost customers is convex for the special cases shown, but it is not known if this is true for the general case.

3. The optimal schedule may assign several customers to arrive at the same time.

4. The optimal loss probabilities are asymmetric between the customers.

5. The $s(1)$-schedule (all customers equally spaced on the interval) is generally not optimal, as opposed to the case where $m = 1$ and $c = 0$.

Theorem 7.2 *If $m = c = 1$ then the optimal schedule is $s(2)$: $\underline{t}^* = (0.5T, 0.5T, 0)$ (i.e., one in the middle of the interval and two at the end) and the optimal loss probabilities are*

$$\begin{aligned}
\ell_1(\underline{t}^*) &= e^{-0.5m\mu T}, \\
\ell_2(\underline{t}^*) &= e^{-m\mu T}\left(1 + 0.5m\mu T\right), \\
\ell_3(\underline{t}^*) &= e^{-m\mu T}\left(e^{0.5m\mu T} + 0.5m\mu T(1 + 0.5m\mu T)\right).
\end{aligned}$$

Remark It is interesting to point out that using $e^{-x}(1+x) < 1,\ \forall x > 0$ we get $\ell_2(\underline{t}^*) < \ell_1(\underline{t}^*) < \ell_3(\underline{t}^*)$.[3]

Theorem 7.3 *If $m \geq 1$ and $c > 1$ the optimal schedule is $s(3)$: $\underline{t}^* = (T, 0, 0)$ (i.e., all three at the end of the interval) and the optimal loss probabilities are*

$$\begin{aligned}
\ell_1(\underline{t}^*) &= e^{-m\mu T}, \\
\ell_2(\underline{t}^*) &= e^{-m\mu T}\left(1 + m\mu T\right), \\
\ell_3(\underline{t}^*) &= e^{-m\mu T}\left(1 + m\mu T + 0.5(m\mu T)^2\right).
\end{aligned}$$

Remark In this case, unsurprisingly we have $\ell_1(\underline{t}^*) < \ell_2(\underline{t}^*) < \ell_3(\underline{t}^*)$.

Proof of Theorems 7.2 and 7.3 Any optimal schedule will assign $m + c$ arrivals at time zero and we are left with scheduling the arrival times of the three remaining customers, denoted by (t_1, t_2, t_3) such that $t_1 + t_2 + t_3 = T$. The blocking probability of the first arrival is simply

$$\ell_1(\underline{t}) = e^{-m\mu t_1}.$$

The second arrival will be blocked if no service completions occurred during $[0, t_1 + t_2)$ or if a single service completion occurred in $[0, t_1)$ and no service completion occurred in $[t_1, t_2)$. Therefore,

$$\ell_2(\underline{t}) = e^{-m\mu(t_1+t_2)}(1 + m\mu t_1).$$

[3]Although two customers try at T, only customer 2 is admitted if there is only one vacant waiting spot, and hence $\ell_2(\underline{t}^*) < \ell_3(\underline{t}^*)$.

The third customer will be blocked if no completions occurred during $(t_1 + t_2, T)$ and one of three following possibilities took place: (1) at most one completion until $t_1 + t_2$, (2) one completion during $[0, t_1)$ and one completion during $[t_1, t_1 + t_2)$, or (3) two completions during $[0, t_1)$ and no completions during $[t_1, t_1 + t_2)$. We state the overall probability for two cases:

$$\ell_3(\underline{t}) = \begin{cases} e^{-m\mu T}\left(e^{m\mu t_1} + m\mu t_2(1 + m\mu t_1)\right), & m = c = 1, \\ e^{-m\mu T}\left(1 + m\mu(t_1 + t_2) + (m\mu)^2(t_1 t_2 + 0.5t_1^2)\right), & m \geq 1,\ c > 1. \end{cases}$$

The overall expected loss we wish to minimize is

$$L(t_1, t_2) := \ell_1(\underline{t}) + \ell_2(\underline{t}) + \ell_3(\underline{t}).$$

It can be verified by computing the corresponding Hessian matrix that $L(t_1, t_2)$ is in fact convex for both of the above cases,[4] although this is not true individually for all components of the sum. Next we observe that the derivative with respect to t_2,

$$\frac{d}{dt_2}L(t_1, t_2) = \begin{cases} m\mu(1 + m\mu t_1)(e^{-m\mu T} - e^{-m\mu(t_1+t_2)}), & m = c = 1 \\ m\mu e^{-m\mu T}\left(1 + m\mu t_1 - e^{m\mu T}\right), & m \geq 1,\ c > 1 \end{cases},$$

is negative for any t_1 in both cases $(1 + x \leq e^x,\ \forall x \geq 0)$. Hence $t_2 = T - t_1$, i.e. T is the optimal arrival time of the second customer. We have thus far established that the two last customers will be assigned at time T. For the assignment of the first customer we consider the first derivative with respect to t_1 at $t_1 + t_2 = T$,

$$\frac{d}{dt_1}L(t_1, t_2)|_{t_1+t_2=T} = m\mu e^{-m\mu T}\left(m\mu(T - 2t_1) + e^{m\mu t_1} - e^{m\mu(T-t_1)}\right),\ m = c = 1,$$

and

$$\frac{d}{dt_1}L(t_1, t_2)|_{t_1+t_2=T} = m\mu e^{-m\mu t_1}\left((m\mu(T - t_1) + 1)e^{-m\mu(T-t_1)} - 1\right),\ m \geq 1,\ c > 1.$$

[4] We have not been able to find an example where the overall objective function is not convex.

In the first case, equating to zero yields a unique solution, $t_1 = 0.5T$. Whereas the second case has no positive solution and is in fact negative for all t_1, which implies that the optimal assignment is $t_1 = T$. We thus get the optimal schedules and loss probabilities as stated in the theorems. ∎

Remark Note that there are other cases that we have omitted, for instance the case of $c = 0$, which would change the blocking probabilities, $\ell_2(\underline{t})$ and $\ell_3(\underline{t})$.

The authors believe that the examples and the insights we have gained from them indicate that finding an optimal schedule in the general case may be a cumbersome task, even numerically. The key open question is whether convexity can be proved for the general case. If the answer is positive then standard numerical descent methods can be applied in order to find the optimal schedule. However, this will still be computationally demanding since the partial derivatives of recursive loss probabilities have no closed form and require numerical analysis as well.

A further constraint on the central planner is having to assign a single symmetric strategy for all customers to use. This may be argued to be more consistent with our claim for the symmetric equilibrium that customers are "anonymous" and cannot coordinate, even when a central planner is involved. Furthermore, the PoA in this case provides an interesting measure of how good or bad the equilibrium arrival profile is, compared with all possible symmetric arrival profiles. This analysis was carried out for waiting cost minimization in [6]. They study a single-server queue where customers can arrive only during an interval $[0, T]$. The optimization is approximated by solving the problem on a discrete grid of the arrival interval, using a local search algorithm. This procedure converges to the global optimum since the objective function is shown to be multimodular (in the sense of [1]). They show that the socially optimal schedule is randomized, has atoms at both zero and T, and a uniform density on all of the interval. We have not been able to show the multi-modularity property for the objective function of minimizing the expected number of losses, and it may not hold for this model. We leave the question of symmetric social optimization open and it is an interesting prospect for future research.

7.1 Numerical examples

In Figure 4 we compare the social utilities achieved by the equilibrium strategy, the $s(k)$-schedule, and the upper bound ($U^e(N)$, $U^{s(k)}(N)$, and $\overline{U}^*(N)$, respectively). We see that for large populations all cases tend towards full utilization. This fact is not surprising since for large values of N both the equilibrium and equally spaced schedules send arrivals at virtually every moment, thus filling an empty waiting spot almost instantly after it is vacated. Clearly, most customers are blocked when N is large under any policy. For small N the customers spread out efficiently in equilibrium and again the expected number of admissions is close to optimal. The $s(k)$-schedule is not far from optimal for all computed N. There is however a medium range of N where the equilibrium is less efficient compared to the other two schedules. This monotone behaviour is better seen in Figure 5, where we compute the bounds for the price of anarchy. This phenomenon also appeared in the explicit computation of the price of anarchy for the single-server case in [21]. The exact choice of $k = 4$ in the example of $m = c = 3$ was the result of the numerical exploration illustrated in Figure 6. The $s(k)$-schedules are better than equilibrium for all k in the example, and there seems to be a concave shape with respect to k with optimal values in the middle of the range $\{1, \ldots, m + c\}$.

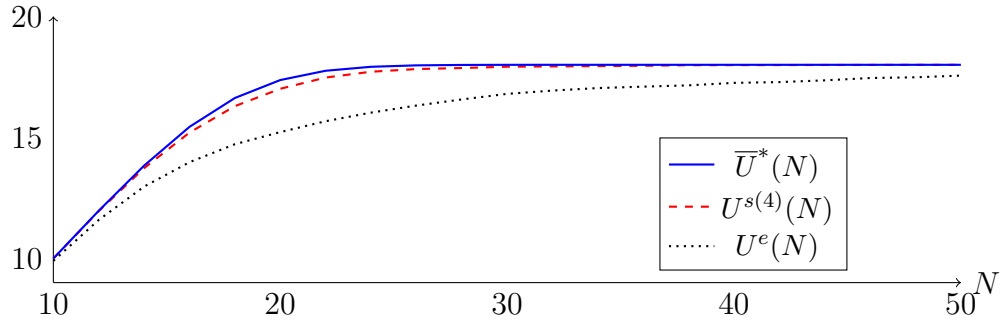

Figure 4: Expected number of customers served for three schemes: equilibrium ($U^e(N)$), equally spaced ($U^{s(4)}(N)$), and dynamic upper bound ($\overline{U}^*(N)$), for different values of N ($\mu = 4, T = 1, m = 3, c = 3$).

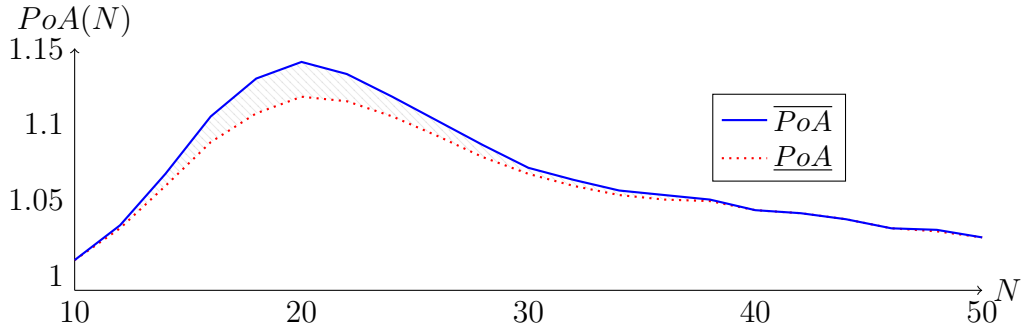

Figure 5: Price of anarchy: lower bound using the equally spaced $(U^{s(1)}(N))$ schedule, and upper bound using the dynamic schedule $(\overline{U}^*(N))$, for different values of N $(\mu = 4, T = 1, m = 3, c = 3)$.

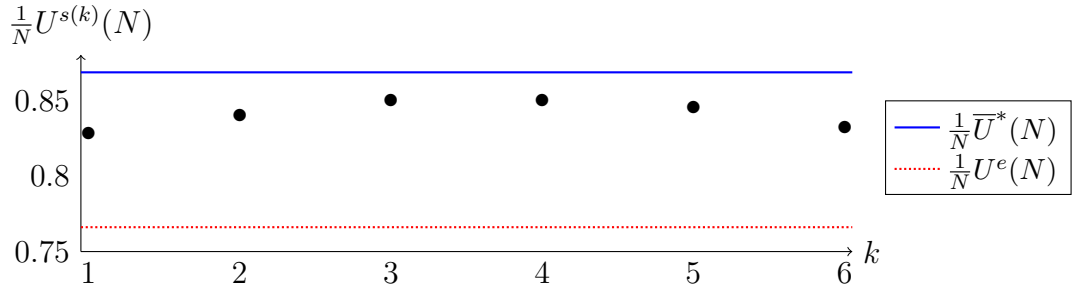

Figure 6: Average admittance probability when using $s(k)$-schedules, for $k = 1, \ldots, m + c$ and parameters $N = 20$, $\mu = 4$, $T = 1$, $m = 3$, and $c = 3$.

8 Conclusion

In this paper we have introduced the $?/M/m/c$ model with strategic arrival times of customers. The focus of the analysis was on minimizing the loss probabilities, both in the game context of selfish customers and in the social optimization context. We have attempted to present general insight despite the difficulties that arise from the combination of a loss queue and the need for transient analysis. We have provided a general characterisation of the symmetric equilibrium arrival distribution, which was shown to be unique, and also described how it can be computed efficiently. Numerical analysis suggests that while the arrival distribution is generally not uniform, it ap-

pears to become so for large populations. This assertion was strengthened by the uniform solution of the fluid approximation model. An interesting extension of the equilibrium analysis is to consider heterogeneous customers, in the sense that the service has a different value for different customers. Another possibility is to combine the loss system with an objective function that includes the value of service and waiting time costs, rather than a binary objective as we have defined it here.

As for the social optimization analysis, we have answered some questions while leaving others open. We nevertheless hope that we have shed some light on the problem, firstly, by providing an upper bound for the price of anarchy and displaying some numerical results that approximate its behaviour, and secondly by discussing the structure of the objective function and the optimal solution for some special cases. There is a lot of room for future research of scheduling arrivals to a loss system, be it in the form of further analysis of the analytical properties of the objective function (e.g., is it convex?) or in the form of efficient approximation algorithms. This may be done for both symmetric and asymmetric schedules.

Acknowledgments

The authors gratefully acknowledge the financial support of Israel Science Foundation grant no. 1319/11.

References

[1] Eitan Altman, Bruno Gaujal, and Arie Hordijk. Multimodularity, convexity, and optimization properties. *Mathematics of Operations Research*, 25(2):324–347, 2000.

[2] William Feller. *An Introduction to Probability Theory and Its Applications*, volume 2. Wiley, 1968.

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

[4] Amihai Glazer and Refael Hassin. Equilibrium arrivals in queues with bulk service at scheduled times. *Transportation Science*, 21(4):273–278, 1987.

[5] Refael Hassin and Moshe Haviv. *To Queue or Not to Queue: Equilibrium behavior in queueing systems*, volume 59. Springer, 2003.

[6] Refael Hassin and Yana Kleiner. Equilibrium and optimal arrival patterns to a server with opening and closing times. *IIE Transactions*, 43(3):164–175, 2011.

[7] Refael Hassin and Sharon Mendel. Scheduling arrivals to queues: A single-server model with no-shows. *Management Science*, 54(3):565–572, 2008.

[8] Moshe Haviv. When to arrive at a queue with tardiness costs? *Performance Evaluation*, 70(6):387–399, 2013.

[9] Moshe Haviv and Igal Milchtaich. Auctions with a random number of identical bidders. *Economics Letters*, 114(2):143–146, 2012.

[10] H. Honnappa and R. Jain. Strategic arrivals into queueing networks. In *Communication, Control, and Computing (Allerton), 2010 48th Annual Allerton Conference on*, pages 820–827, 2010.

[11] Harsha Honnappa, Rahul Jain, and Amy R. Ward. $\Delta(i)$ /GI/ 1: A New Queueing Model For Transitory Queueing Systems. arXiv:1206.0720, 2012.

[12] Rahul Jain, Sandeep Juneja, and Nahum Shimkin. The concert queueing game: To wait or to be late. *Discrete Event Dynamic Systems*, 21(1):103–138, 2011.

[13] S. Juneja and T. Raheja. The concert queueing game: Fluid regime with random order service. *International Game Theory Review (forthcoming)*.

[14] Sandeep Juneja and Nahum Shimkin. The concert queueing game: strategic arrivals with waiting and tardiness costs. *Queueing Systems*, 74(4):369–402, 2013.

[15] Guido C. Kaandorp and Ger Koole. Optimal outpatient appointment scheduling. *Health Care Management Science*, 10:217–229, 2007.

[16] Kyle Y. Lin and Sheldon M. Ross. Optimal admission control for a single-server loss queue. *Journal of Applied Probability*, 41(2):535–546, 2004.

[17] V. V. Mazalov and J. V. Chuiko. Nash equilibrium in the optimal arrival time problem. *Computational Technologies*, 11:60–71, 2006.

[18] Hironori Otsubo and Amnon Rapoport. Vickreys model of traffic congestion discretized. *Transportation Research Part B: Methodological*, 42(10):873–889, 2008.

[19] Claude Dennis Pegden and Matthew Rosenshine. Scheduling arrivals to queues. *Computers & Operations Research*, 17(4):343–348, 1990.

[20] Liron Ravner. Equilibrium arrival times to a queue with order penalties. *European Journal of Operational Research*, 239(2):456–468, 2014.

[21] Liron Ravner and Moshe Haviv. Equilibrium and socially optimal arrivals to a single server loss system. In *International Conference on NETwork Games COntrol and OPtimization 2014 (NetGCoop'14)*, Trento, Italy, October 2014.

[22] William E. Stein and Murray J. Côté. Scheduling arrivals to a queue. *Computers & Operations Research*, 21(6):607–614, July 1994.

[23] William S. Vickrey. Congestion theory and transport investment. *The American Economic Review*, 59(2):251–260, 1969.