Equilibrium and optimal arrival patterns to a server with opening and closing times

Refael Hassin and [1] and Yana Kleiner [2]

Abstract

We consider a first-come first-served single-server system with opening and closing times. Service durations are exponentially distributed, and the total number of arrivals is a Poisson random variable. Naturally each customer wishes to minimize his waiting time. The process of choosing an arrival time is presented as a (non-cooperative) multiplayer game. Our goal is to find a Nash equilibrium game strategy. Glazer and Hassin (1983) assume that arrivals before the opening time of the system are allowed. We study the case where early arrivals are forbidden. It turns out that unless the system is very heavily loaded, the equilibrium solution with such a restriction does not reduce the expected waiting time in a significant way. We also compare the equilibrium solution with the solution which maximizes social welfare. Finally, we show how social welfare can be increased in equilibrium by restricting arrivals to certain points of time.

Keywords: Queues: Markovian, Non-stationary, Equilibrium Arrivals.

1 Introduction

Many real life service systems work in noncontinuous manner, having opening and closing times. Post office services, banks, government offices, libraries, shops, drugstores, and bag drop-off services supplied by airlines are just a few of these services. However, with only few exceptions, existing theory assumes that service systems are continuously open to accept new arrivals.

A notable exception is the paper by Glazer and Hassin (1983); we adopt their assumptions in this paper. The model considers a single-server system, which opens at time zero, closes at time T, and applies a first-come first-served discipline. All customers who arrive before closing time are admitted to the queue, and the server is available to complete their service as necessary. Customers choose their arrival times independently. Customers are indifferent to the exact time of the day they spend in the system, and their goal is to minimize their expected length of waiting time.

A central assumption in our model is that customers are free to choose their arrival time and have no preference to their exact time of service during

[1]School of Mathematical Sciences, Tel Aviv University, Tel Aviv 69978, Israel, hassin@post.tau.ac.il Research supported by Israel Science Foundation Grant no. 526/08.

[2]School of Mathematical Sciences, Tel Aviv University, Tel Aviv 69978, Israel, yanaklei@post.tau.ac.il

the day. We discuss this *full flexibility assumption* and its implications in Section 3.3.

The waiting time of a customer is affected by the decisions of the other customers, and this turns the process into a non-cooperative multi-player game. A strategy in this game is expressed by a density function giving the probability distribution of arrival time for a random customer. The equilibrium strategy, in general, is not optimal, in the sense that it doesn't minimize the expected total waiting time of all customers. The reason is that when deciding when to use the system, a user does not take into account the effect of his decision on the waiting time of other users (which is referred to as the external cost).

Assuming exponential service distribution and that the total number of arrivals is a Poisson random variable, Glazer and Hassin compute the equilibrium arrival density. An important feature of this model is that early arrivals, that is, arrivals before opening time, are allowed.

A common way to reduce waiting time is by setting appointments. The use of appointments has been studied extensively, see for example Cayirli and Veral (2003). Such systems are commonly seen in health-care services, but not used by banking systems, for example.

In this paper we approximate the optimal density function, which minimizes the expected waiting time over all arriving customers (it is not an equilibrium solution and therefore arrivals at different instants are associated with different expected waiting time).

Since one usually cannot force customers to behave in a non-equilibrium manner, this solution cannot be implemented. However, one might still try to reduce the waiting time by restricting the instants when arrivals are admitted. We consider two types of such restrictions. In the first, early arrivals are eliminated (for example by randomly ordering the customers at time 0). In the second, new arrivals are admitted only at a small set of instants. We show that such restrictions often make it possible to obtain an equilibrium with reduced expected waiting time.

The following illustrative example is a variation of the model, and the insights we obtain can also help in finding ways to deal with such instances. The port of Haifa is located in the north of Israel. Trucks line up to obtain a container which they transport in most cases to the center of Israel, and then return for another trip. Containers are released from 6am, at which time there is a very long line of trucks whose drivers arrived earlier to secure a position in the line. After the first delivery of the day, trucks normally return in the afternoon for another pickup. It is reasonable to assume that drivers are sufficiently flexible with respect to when they start their working day, and that their main goal is to obtain a container at minimum wait as long as it enables them to return for a second round. Is it possible to reduce the drivers waiting time by applying a random service order for all those

present at 6am?

We show that eliminating early arrivals reduces the expected waiting time significantly only when the system's load is heavy (equivalently, the opening interval is short). However, when the server's utilization is not very low such a change will hardly save any waiting time. On the other hand, we show how reduction in waiting can be achieved by other simple restrictions on the admission times of customers.

The paper is structured as follows. In Section 2 we survey relevant literature. The queueing model is presented and the equilibrium is computed in Section 3. The optimal solution is studied in Section 4. In Section 5 we present numerical results and compare the models, and in Section 6 we analyze restrictions of the arrivals to small sets of points. Finally, in Section 7 we summarize our main results.

2 Literature Review

The model discussed in this paper is new and has not been studied in this specific form. However, it draws many similarities to previous results. In this paper we mention related models and discuss their relevance.

2.1 Transportation models

Research concerning equilibrium timing decisions was initiated in the context of transportation models. Vickrey (1969) described a deterministic model where commuters have complete knowledge about the other commuters choice, and can change their home departure time to avoid periods of high congestion. This simple model assumes a single bottleneck and a fixed number of commuters. Delay occurs when the traffic flow exceeds the capacity of the bottleneck. Each commuter has a preferred time to pass the bottleneck. There is a cost for arriving earlier than desired, and a cost for a late arrival. For later developments on this model see Lago and Daganzo (2007), Arnott (1999), Ostubo and Rapoport (2007), and their bibliographies. A major difference between these models and our queueing model is that we assume that customers do not have time preference and the equilibrium we compute is symmetric. Moreover, congestion cannot be eliminated (though it can be alleviated) even when customers behave in a socially optimal way.

2.2 Queueing models

The first queueing model where the arrival process is endogenously determined is Glazer and Hassin (1983), as already described in the introduction.

Rapoport, Stein, Parco and Seale (2004) (see also, Bearden, Rapoport and Seale (2005)), considered a closely related discrete-time model of a

queueing system with pre-specified and commonly known opening and closing times, a fixed and commonly known number of players, fixed service time, and no early arrivals. A player who arrives late and his service cannot be completed before closing time is not served. A companion paper, Seale et al. (2005), studies the same queueing system but allowing early arrivals. The papers compute equilibrium solutions and verify them experimentally. The experimental results indicate convergence to equilibrium with experience in playing the game.

Glazer and Hassin (1987) compute equilibrium arrivals to a server with batch service scheduled at published instants separated by intervals of a constant length. Customers who arrive close to the beginning of service may face a full batch and need to wait for the next service time. This model has no opening and closing times, and what makes it non-stationary is the common knowledge of the fixed schedule.

Lariviere and Van Mieghem (2004) assume that each customer chooses a day within a given interval to arrive at a server. The waiting cost is monotone increasing with the number of customers arriving on a particular day. Ideally, customers should split as evenly as possible over the time interval, and this is also an equilibrium if they have information on each other's decision. When this information is not available, in a symmetric equilibrium, each customer uniformly chooses a day for arrival. In this case, when the number of customers grows, the arrival distribution approaches Poisson. Lariviere and Van Mieghem also consider a capacitated version of their model which is more relevant to our discussion. In this version customers not served in period t carry over to period $t+1$. This version resembles the model of Glazer and Hassin (1987).

Mazalov and Chuiko (2006) consider a single server with no queue buffer. There is a "convenience" function $C(t)$, that expresses a desirability of a service starting at time t, that a player receives if his request arrives at time t and is served successfully (i.e., the server is not engaged in serving another request). A player's strategy is a distribution density of arrival time t.

Wang and Zhu (2004) consider demand that is processed in shifts. Every customer chooses one of a set of shifts. There is a different cost for waiting till the shift and for waiting during the shift till the demand is processed. All customers prefer service at an early time, but in equilibrium the expected wait in early shifts is higher, and therefore they distribute their demand over the shifts.

Guo, Liu and Wang (2009) consider equilibrium behavior in a single-server two-period queueing model where customers decide in the first period when to arrive to the queue based on their current information and anticipated future gains.

Queueing models with customers' arrival-time decisions are the subject of Chapter 6 of Hassin and Haviv (2003). Other recent non-queueing models

with arrival-time decisions in equilibrium are Arbatskaya, Mukhopadhaya and Rasmusen (2007) and Ostrovsky and Schwarz (2005, 2006).

3 Equilibrium arrival pattern

3.1 The model

The system consists of a single server open for arrivals each day during a given time interval $[0, T]$. The server is available to serve all who arrive during this interval, including those who remain in the system after time T. As mentioned in the introduction, the system with early arrivals has been solved in Glazer and Hassin (1983). We now treat the model in which arrivals are restricted to $[0, T]$.

The service duration of a customer is distributed exponentially with parameter μ. The number of customer arrivals *in a day* is a Poisson random variable with parameter λ. Note that the Poisson distribution is obtained when each individual from a large population independently decides at any given day whether he needs service during that day (i.e., the Poisson distribution here is the limit of the binomial distribution). Also note that there is no need to require $\lambda < \mu T$ for stability of the model, because the server is available to continue service as necessary until all customers who arrived during the day are served.

Each arriving customer independently chooses his arrival time, with the objective of minimizing his expected waiting time. Since each customer's decision affects the waiting time of other customers, customers take into consideration the strategies of the others. Hence, we consider (Nash) equilibrium strategies: If we observe that a customer arrives at time t_1, and another arrives at time t_2, then it must be that the expected waiting time is identical at these two points, and moreover, it is not greater than the expected waiting time at any other instant in $[0, T]$.

Let λ be the expected number of customer arrivals during the day. Let $N(t)$ be the number of customers in the system at time t, and let μ be the service rate.

We compute symmetric solutions, assuming that customers draw their arrival time using a common density function $f(t)$ $t \in [0, T]$. Since by assumption, the total number of arrivals during the day is Poisson, the equilibrium arrival process is non-homogeneous Poisson, and the numbers of arrivals in disjoint intervals are independent random variables (see Kingman (1993) §4.5 for a proof in a more general context). In particular, the rate of arrivals at t, $\lambda(t) = \lambda f(t)$, is independent of the arrivals history, in particular of $N(t)$.

Customers who arrive simultaneously are served in random order. We will see that the opening time is the only time where multiple arrivals are

possible.

3.2 The solution

Let w be the equilibrium expected waiting time. Note that in this paper the waiting time is the time a customer spends in the queue, not including service time.

An equilibrium solution has the following properties:

- There is a positive probability of arrival at time 0. Otherwise arrival at time 0 guarantees zero wait, contradicting the equilibrium conditions.
- There is an open time interval starting at 0, say $(0, t')$, where there are no arrivals. The reason is that for $t > 0$ sufficiently small, arriving at time 0 rather than at t would decrease the expected waiting time. In fact, by this change the customer saves in expectation the wait for half of the customers who arrive at time 0.
- There is no positive probability of arrival at any $t > 0$. If there was such an instant then arriving just before it would guarantee a strictly shorter expected waiting time.

Thus, the solution is characterized by the probability p_0 to arrive at time 0, and a continuous density function $f(t)$ $t \in (0, T]$ which determines the arrival density after opening. We divide $[0, T]$ into three parts: the point 0, the interval $(0, t']$, and the interval $(t', T]$.

$t = 0$. The point 0 is special. There is a positive probability p_0, for the customers to arrive at time 0. The expected number of customers arriving at time 0 is therefore λp_0, and a customer who decides to arrive at time 0 will have on the average to wait for half of these customers. Hence, the expected waiting time of customers arriving at 0 is $\frac{\lambda p_0}{2\mu}$.

In equilibrium, the expected waiting time for every customer is w. In particular, the expected waiting time for a customer arriving at 0 is w. So, $\frac{\lambda p_0}{2\mu} = w$, and

$$p_0 = \frac{2\mu w}{\lambda}. \tag{1}$$

Note that if all customers arrived at the same instant then the expected waiting time of a customer would be $\frac{\lambda}{2\mu}$, and this is an upper bound on w. Therefore, p_0 in (1) is in [0,1].

$0 < t < t'$. Suppose that a customer arrives immediately after time 0. Then, his expected waiting time is $\frac{\lambda p_0}{\mu} = 2w$. Clearly, the density function has

to vanish from 0 to some point t', such that if the new customer arrives to the system at time t' then his expected waiting time would be w. In other words, $f(t) = 0$ for all $0 < t < t'$ and $E[N(t')] = \mu w$.

$E[N(0)] = 2\mu w$, and as long as the server is busy, the expected number of customers in the system decreases at rate μ per time unit. This means that if the server were guaranteed to be busy continuously during $[0, t]$ we would have $E[N(t)] = E[N(0)] - \mu t = \mu(2w - t)$ and with $t' = w$ we would have $E[N(t')] = \mu w$, so that the expected wait of a new arrival at time w is w, and from this point we would have $f(t) > 0$. However, there is a positive probability that all customers who arrive at $t = 0$ are served before time $t = w$, and hence $E[N(t)]$ decreases at a lower rate than μ, giving that

$$t' > w. \tag{2}$$

$t' \le t \le T$. The expected waiting time for a customer who arrives at t is $E[N(t)]/\mu$. (Note that we refer to the expected number of customers at t given the equilibrium arrival process, and not to the overall expected number of customers at a random instant, and therefore we do not need to claim that the arrivals see time averages (PASTA).) We conclude that for a customer's expected waiting time to be constant in $[t', T]$, also $E[N(t)]$ must be constant in this interval.

Let $P_k(t)$ be the probability that exactly k persons are in the system at time t. Then,

$$E[N(t + dt)] = E[N(t)] + \lambda f(t)dt - \mu(1 - P_0(t))dt.$$

Applying the equilibrium's condition $E[N(t + dt)] = E[N(t)]$, we obtain

$$f(t) = [1 - P_0(t)]\frac{\mu}{\lambda}.$$

Altogether we have the following equations:

$$p_0 = \frac{2\mu w}{\lambda}, \tag{3}$$

$$P_k(0) \sim \text{Poisson}(2w\mu). \tag{4}$$

For $0 \le t \le t'$:

$$f(t) = 0, \tag{5}$$

$$P_0'(t) = P_1(t)\mu, \tag{6}$$

$$P_k'(t) = \mu\left[P_{k+1}(t) - P_k(t)\right], \qquad k = 1, 2, \ldots \tag{7}$$

The value of t' is determined by

$$t' = \min\left\{T, \inf\left\{t \ : \ \sum_{k=1}^{\infty} k \cdot P_k(t) = \mu w\right\}\right\}. \tag{8}$$

For $t > t'$:

$$f(t) = [1 - P_0(t)]\frac{\mu}{\lambda}, \tag{9}$$

$$P_0'(t) = P_1(t)\mu - P_0(t)\lambda f(t), \tag{10}$$

$$P_k'(t) = P_{k-1}(t)\lambda f(t) + P_{k+1}(t)\mu - P_k(t)[\lambda f(t) + \mu], \qquad k = 1, 2, \ldots \tag{11}$$

For a given value of w we use the above formulas to compute $f(t)$ $0 \leq t \leq T$. Since the value of w is unknown, we search for a value such that the resulting density function satisfies

$$p_0 + \int_{t'}^{T} f(t)dt + p_T = 1. \tag{12}$$

3.3 The full flexibility assumption

The assumption that customers have no preference to the time of service is central in our model. There are many situations where this assumption cannot be made, such as patients waiting for emergency medical treatment or drivers on their way to work. However, there are numerous situations where customers are more flexible in choosing their arrival time, for example when taking one's car for an emission inspection, visiting the bank or the post office, visiting the drugstore, etc. A similar decision is faced by a customer making a telephone call to obtain information or order a service while trying to avoid a long wait.

Our results are applicable in more general contexts where the population consists of a mixture of *flexible* and *inflexible* customers. The inflexible customers arrival times are exogenously fixed. Their expected number during the day is λ_1 and the arrival process they generate has probability q_0 of arrival at opening time and a density $g(t)$ $t \in (0, T]$ afterwards. The flexible customers are free to choose their arrival-time, and they choose it so as to minimize their expected waiting time. Their expected number is λ_2. These customers take into consideration the existence of inflexible customers.

Such a mixed model is reflected in the experimental outcome of Rapoport et. al (2004) who observed that "some players stick to the same arrival time and others switch their arrival times across rounds in an attempt to increase their payoffs." We can think of the first type of players as inflexible customers and the second type as flexible. In the experiments, in spite of the existence of players of the first type, eventually in the equilibrium all players had the same expected waiting time. Facing a mixed situation, and trying to expedite convergence to an equilibrium, some servers announce

recommended intervals for arrival, trying to induce its flexible customers to avoid hours that are much used by its inflexible customers. For example, the State of Delaware Division of Motor Vehicles recently announced that "normally, the best times for short waits are between 8:00-11:00 a.m. and 2:00-4:00 p.m. on Tuesdays, Thursdays and Fridays when these days do not fall close to a registration expiration period."

Let p_0 and $f(t)$ $t \in (0, T]$ be the equilibrium solution obtained under the full flexibility assumption with $\lambda = \lambda_1 + \lambda_2$. We observe that if $\lambda_1 q_0 \leq \lambda p_0$ and $\lambda_1 g(t) \leq \lambda f(t)$ for $t \in [0, T]$ then the equilibrium solution will be exactly as under the full flexibility assumption, and the expected waiting time of all customers, flexible or not, will be identical. Therefore, our results apply as long as the inflexible customers do not overload the system at any point of time relative to the computed equilibrium.

Suppose now that the arrival rate induced by inflexible customers exceeds the rate under the flexibility assumption. In this case a different equilibrium will be obtained, in which flexible customers refrain from arriving at certain time intervals where the expected waiting caused by the workload induced by inflexible customers already exceeds the equilibrium expected waiting time for flexible customers. The resulting equilibrium can be computed in a similar way to the one followed in this section, for any given input q_0 and $g(t)$.

4 Socially optimal solution

In the previous section we computed the equilibrium density function. In this section we compute an approximation to the density function that minimizes the expected waiting time.

For numerical computation we applied twofold discretization: We define two small positive quantities Δ_t and Δ_p. Customers' arrivals are restricted to discrete points $0, \Delta_t, 2\Delta_t, \ldots, T$. For every $t = 0, \Delta_t, \ldots, T$ let p_t denote the probability of arrival at time t, with the obvious condition $\sum_{t=0}^{T} p_t = 1$. We further restrict the probabilities p_t to be integer multiples of Δ_p. The resulting model becomes exactly the outpatient scheduling problem solved by Kaandorp and Koole (2007) where each probability quantum Δ_p represents a customer, and customers need to be scheduled to arrive at $0, \Delta_t, \ldots, 1$.

Kaandorp and Koole (2007) prove that their objective function is *multimodular*. Multimodularity was introduced by Hajek (1985) and extended by Altman, Gaujal and Hordijk (2000). It is a property of functions on a lattice, related to convexity. Koole and van der Sluis (2003) prove that for a system under certain conditions, local search applied on a limited neighborhood provides a global minimum. The crucial condition is multimodularity of the objective function.

We applied local search to compute an optimal density function, starting from a trial vector $p = (p_0, p_{\Delta_t}, \ldots, p_T)$ and striving to reduce the expected waiting time.

The size of the neighborhood required by a direct application of Koole and van der Sluis (2003) is limited but still exponential in the number of points $\frac{T}{\Delta}$. However, satisfying results in practice can also be obtained by considering only a subset of this neighborhood.

In each iteration we compute the expected waiting time for the given probability vector $p = (p_0, p_{\Delta_t}, \ldots, p_T)$. First we calculate the probabilities $P_k(t)$ that exactly k customers are in the system at time t. Let X_t be a Poisson random variable with parameter λp_t, representing the number of arrivals at discrete point t. The probabilities $P_k(t)$ can be calculated using the following recursive equations:

$$P_k(0) \sim \text{Poisson}(\lambda p_0), \tag{13}$$

$$P_0(t + \Delta_t) = P_0(t)P[X_{t+\Delta_t} = 0] + P_1(t)P[X_{t+\Delta_t} = 0]\mu\Delta_t, \tag{14}$$

$$P_k(t+\Delta_t) = \sum_{i=0}^{k} P_i(t)P[X_{t+\Delta_t} = k-i](1-\mu\Delta_t)+\sum_{i=0}^{k+1} P_i(t)P[X_{t+\Delta_t} = k-i+1]\mu\Delta_t. \tag{15}$$

In equations (14) and (15) we assume that at most one service is completed in the interval of length Δ_t, since Δ_t is small and the service time is distributed exponentially.

Using Equations (13)-(15) we calculate the expected waiting time E_t for a customer who arrives to the system at time t:

$$E_0 = \frac{\lambda p_0}{2\mu} \tag{16}$$

$$E_t = \sum_{k=0}^{N} P_k(t - \Delta_t)(1 - \mu\Delta_t)\frac{k}{\mu} + \sum_{k=0}^{N} P_{k+1}(t - \Delta_t)\mu\Delta_t\frac{k}{\mu} + \frac{\lambda p_t}{2\mu}, \tag{17}$$

where $t = 0, \Delta_t, \ldots, T$ and N is a number large enough such that the probability that there are more than N customers in the system at any time t is negligible. The term $\frac{\lambda p_t}{2\mu}$ in these equations arises due to the following consideration: all customers arriving at the same time are served in random order. Thus, on the average a customer will have to wait for half of the customers who arrive with him to be served.

Finally, the expected waiting time is calculated as $\sum_t E_t p_t$.

We computed the solution for various values of λ and μ. All density functions obtained have the same property: The density function is approximately uniform in $(0, T)$, and there are positive probabilities p_0 and p_T, for arrivals at time 0 and T, respectively. For example, in Figure 1 one can see a typical density function. It gives the density function on $(0, T)$ and also depicts the probabilities p_0 and p_T.

μ	$\lambda = 10$		$\lambda = 15$		$\lambda = 20$	
	w_{opt}	w_{apx}	w_{opt}	w_{apx}	w_{opt}	w_{apx}
8	.238	.239	.439	.443	.673	.681
10	.154	.155	.292	.295	.461	.466
12	.105	.106	.203	.205	.329	.332
14	.069	.075	.147	.147	.242	.244
15	.064	.064	.126	.126	.209	.230
16	.056	.056	.109	.109	.181	.183
18	.042	.042	.082	.083	.139	.140
20	.033	.033	.064	.064	.108	.108
30	.012	.012	.023	.023	.038	.038

Table 1: Minimum expected waiting time w_{opt} vs. the expected waiting time w_{apx} under the approximate optimum

Consequently, we found that it is possible to assume with almost no loss of accuracy that the density function which minimizes the expected waiting time has this form. The desired function depends only on two parameters, p_0 and p_T. This way the computations can be simplified. We refer to the resulting solution as the *approximate optimum*.

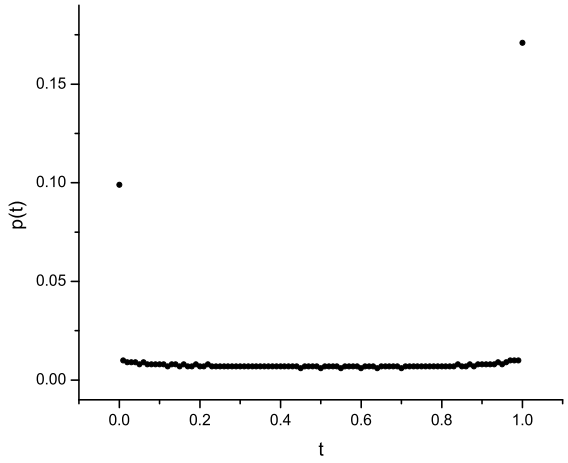


Figure 1: Optimal density function

Table 1 compares the expected waiting time under the optimal solution and the approximate optimum.

μ	$\lambda = 10$	$\lambda = 12$	$\lambda = 15$	$\lambda = 20$
8	.405	.583	.902	1.500
10	.238	.348	.562	1.009
12	.151	.220	.362	.691
14	.101	.146	.241	.479
15	.085	.121	.199	.403
16	.072	.102	.166	.336
18	.053	.074	.119	.240
20	.041	.056	.088	.174
30	.015	.020	.029	.050

Table 2: Expected waiting time in equilibrium with early arrivals

5 Numerical Results

In this section we present some numerical computations for the models described above.

We first compute the equilibrium density function with early arrivals, as in Glazer and Hassin (1983). Table 2 presents the equilibrium expected waiting time w for various values of λ and μ. Density functions that bring the system to equilibrium are presented in Figure 2 (left).

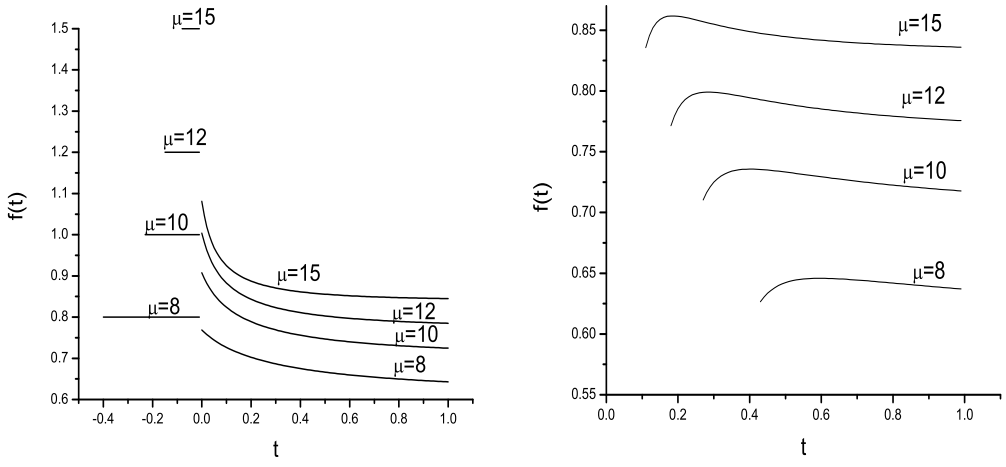


Figure 2: Equilibrium density functions with (left) and without (right) early arrivals ($\lambda = 10$)

μ	$\lambda = 10$			$\lambda = 15$			$\lambda = 20$		
	w	t'	p_0	w	t'	p_0	w	t'	p_0
8	.397	.43	.635	.895	.92	.955	1.250	-	1
10	.231	.27	.470	.555	.58	.740	1.000	-	1
12	.148	.18	.355	.355	.38	.568	.690	.72	.835
14	.100	.14	.280	.238	.26	.439	.478	.50	.669
15	.083	.11	.249	.198	.23	.396	.399	.43	.605
16	.068	.09	.218	.166	.20	.354	.331	.35	.530
18	.050	.07	.180	.118	.15	.283	.238	.26	.428
20	.039	.06	.156	.088	.12	.235	.170	.19	.340
30	.012	.02	.072	.027	.05	.108	.049	.06	.147

Table 3: Equilibrium solutions without early arrivals

μ	$\lambda = 10$			$\lambda = 15$			$\lambda = 20$		
	w	p_0	p_T	w	p_0	p_T	w	p_0	p_T
8	.238	.082	.299	.443	.006	.391	.681	.048	.464
10	.154	.082	.248	.294	.006	.337	.466	.051	.412
12	.105	.081	.206	.205	.007	.289	.332	.053	.364
14	.069	.077	.172	.147	.007	.248	.244	.056	.321
15	.064	.075	.158	.126	.007	.230	.210	.057	.301
16	.059	.073	.146	.109	.007	.213	.183	.058	.282
18	.052	.069	.125	.083	.008	.183	.139	.059	.248
20	.033	.064	.108	.064	.008	.157	.108	.060	.217
30	.012	.044	.062	.023	.009	.080	.038	.051	.112

Table 4: The approximate optimum

Figure 2 (right) depicts equilibrium density functions and Table 3 presents the expected waiting time w for various values of λ and μ when early arrivals are forbidden. For each μ there are given p_0, w, and t'.

Recall from Equation (2) that $t' \geq w$. In Table 3 we compare the values of w and t'. Note that when λ is large and μ is small, all customers arrive at the time of opening, i.e., $p_0 = 1$.

Table 4 presents the approximate optimum (p_0, p_T), and the corresponding value of w for various values of μ and λ.

μ	$\lambda = 10$			$\lambda = 15$			$\lambda = 20$		
	w_1	w_2	w_{apx}	w_1	w_2	w_{apx}	w_1	w_2	w_{apx}
8	.405	.397	.239	.902	.895	.443	1.500	1.250	.681
10	.238	.231	.154	.562	.555	.294	1.009	1.000	.466
12	.151	.148	.105	.362	.355	.205	.691	.690	.332
14	.101	.100	.075	.241	.238	.147	.479	.478	.244
15	.085	.083	.064	.199	.198	.126	.403	.399	.230
16	.072	.068	.056	.166	.166	.109	.336	.331	.183
18	.053	.050	.042	.119	.118	.083	.240	.238	.140
20	.041	.039	.033	.088	.088	.064	.174	.170	.108
30	.015	.012	.012	.029	.027	.023	.050	.049	.038

Table 5: Expected waiting time: w_1 refers to equilibrium with early arrivals; w_2 to equilibrium without early arrivals; w_{apx} to the approximate optimum

Table 5 compares the expected waiting time in the three models. We see that the suboptimality of the equilibrium is greater when λ is large and μ is small. For small λ and big μ the difference is very small.

Figure 3 presents expected waiting times for the three models. We see that when the system is not heavily loaded, the expected waiting times for the two models of the system in equilibrium are very similar, despite the difference in the arrival density functions, as shown in Figure 4.

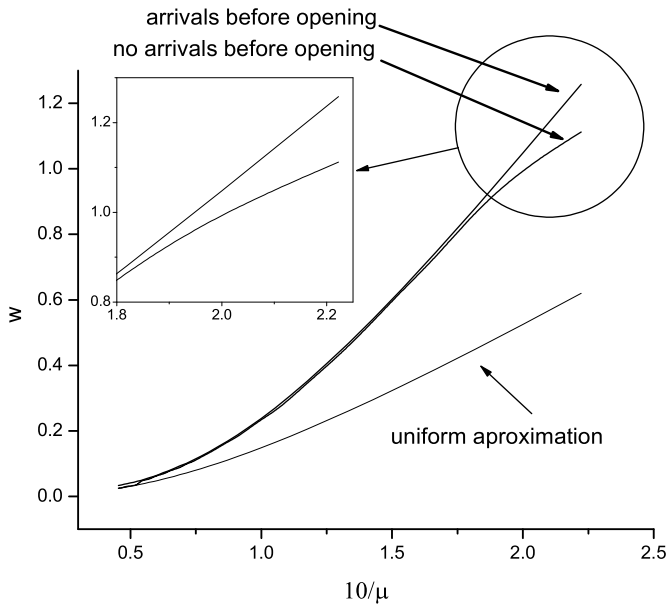


Figure 3: A comparison of the three models. The results are presented for $\lambda = 10$.

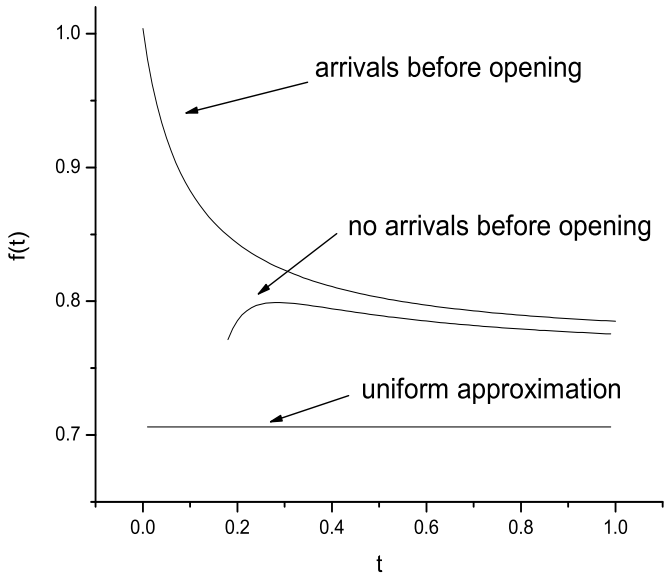
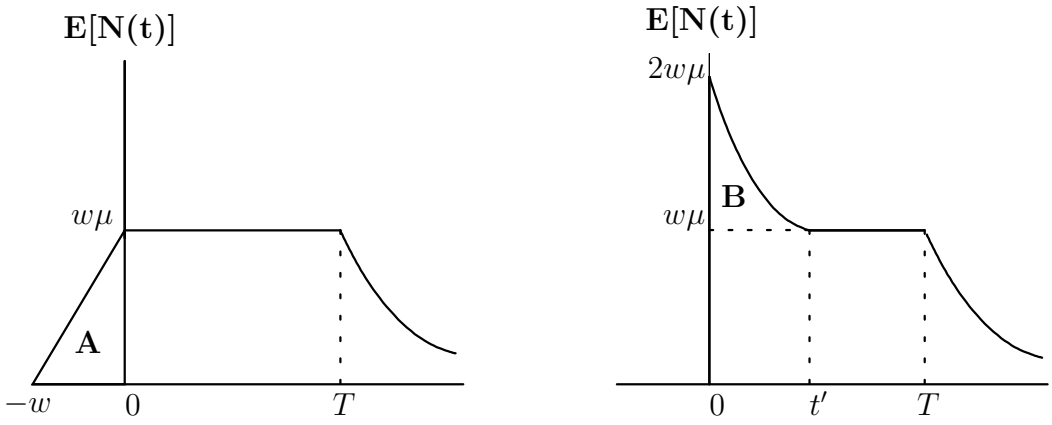


Figure 4: Density functions ($\lambda = 10$, $\mu = 10$): equilibrium with early arrivals; equilibrium when early arrivals are forbidden; approximate optimum (The graphs do not integrate to 1 because of positive probabilities at 0 and 1)

For an explanation of this result consider Figure 5. These graphs describe $E[N(t)]$, the equilibrium expected number of customers at time t, when early arrivals are allowed (Figure 5(a)) and forbidden (Figure 5(b)).



(a) Arrivals before opening (b) No arrivals before opening

Figure 5: The expected number of customers

There are three intervals in Figure 5(a). In the first, from $-w$ to 0, $f(t) = \frac{\mu}{\lambda}$, and $E[N(t)]$ increases linearly from 0 to $w\mu$. From 0 to T the expected number of customers is constant and equals $w\mu$, and for $t > T$ it decreases.

There are also three intervals in Figure 5(b). The expected number of arrivals at time 0 is $2w\mu$. From 0 to t', the number of customers decreases to $w\mu$. Between t' and T $E[N(t)]$ stays equal to $w\mu$, and then, for $t > T$ it decreases, as in the first model.

The area under $E[N(t)]$ is proportional to the expected waiting time. The effect of eliminating early arrivals is twofold. On the one hand, customers who would otherwise arrive before opening now arrive at $t = 0$, saving the wait represented by area **A** in Figure 5(a). On the other hand, customers who would otherwise arrive in $(0, t')$ also arrive at the opening time, and this adds area **B** in Figure 5(b). These changes have opposite effects on the total (or expected) waiting time which, and when the system is not extremely overloaded, they approximately cancel each other.

Consider again Figure 3. When the system is highly loaded (small μ or large λ), the expected waiting time when early arrivals are forbidden becomes significantly smaller than when early arrivals are allowed. To understand this behavior, consider the case with very large $\frac{\lambda}{\mu}$. This is equivalent to assuming that T is very small. Consider the extreme case with $T = 0$. In this case it is obvious that excluding early arrivals is desirable. In fact, when early arrivals are allowed, w is the expected waiting time of a customer who

arrives at $t = 0$ and is therefore guaranteed to be the last one. He will wait for all of the other customers to be served so $w = \frac{\lambda}{\mu}$. If, on the other hand, early arrivals are not allowed, all come at time 0 and $w = \frac{\lambda}{2\mu}$. Thus, in the limit, excluding early arrivals saves half of the waiting time. However, as we see from the figure, with $\lambda = 10$, excluding early arrivals has almost no effect for $\mu > 5$. Note that $\mu \leq 5$ means that the server needs at least twice the length of time the system is open in order to serve the demand – quite an uncommon situation.

6 Discrete points

We have seen that the expected waiting time of the equilibrium solution, even when early arrivals are forbidden, may be much higher than under an optimal solution. Clearly, in most cases it is not possible to induce customers to cooperate and behave in the optimal way. In this section we show that by restricting the time intervals in which the system admits new arrivals, it is possible to obtain arrival density functions that better resemble the optimal one, and in this way reduce the expected waiting time.

These "degenerate" solutions cannot be part of a first-best socially optimal solution, but customers do not follow a socially optimal behavior and we should be satisfied with second-best solutions. It turns out that by restricting the arrivals we can obtain better solutions than the unrestricted equilibrium.

We consider two simple models that are easy to implement.

6.1 Two points

Suppose that we restrict admission to two points of time. The optimal choice for these points is clearly 0 and T. The question is whether it is possible in this way to reduce the expected waiting time relative to the original model.

In equilibrium, either all customers arrive at time 0, or the expected waiting time of customers arriving at times 0 and T are equal. Consider the latter case. The expected number of customers arriving at 0 is λp_0, and their expected waiting time is therefore $\frac{p_0\lambda}{2\mu}$. Denote by $R > 0$ the expected remaining time of service after time T, of customers who arrived at time 0. Then, taking into account that the expected number of customers who arrive at T is $\lambda(1-p_0)$, the expected waiting time of a customer who arrives at T is $R + \frac{(1-p_0)\lambda}{2\mu}$. In equilibrium,

$$w = \frac{p_0\lambda}{2\mu} = R + \frac{(1-p_0)\lambda}{2\mu}.$$

Since $R > 0$,

$$\frac{p_0\lambda}{2\mu} > \frac{(1-p_0)\lambda}{2\mu} \Longrightarrow p_0 > 0.5.$$

Table 6 shows some numerical results comparing the equilibrium solutions when customers are admitted continuously in $[0, T]$ (in the two columns marked *Continuous*) and when they are only admitted at 0 and T.

The two-points solution reduces the expected waiting time when the system is highly utilized, but not in the extreme cases where $p_0 = 1$ under equilibrium in the continuous model, because in this case we clearly also obtain $p_0 = 1$ in the two-points equilibrium.

6.2 Three points

Trying to further reduce the expected waiting time, we allow customer arrivals in three discrete points. Clearly, the optimal choice contains 0,T and some intermediate point which we will choose optimally. Some results are presented in Table 7. The expected waiting time in this model is clearly less then in the two-points model, since the two-points model is a special case with two of the three points located together. For every three-points solution we give the equilibrium expected waiting time w, the probabilities p_0 and p_T of arriving at 0 and T, respectively, and the middle point t_m.

Tables 8 and 9 compare the four models: approximately optimal solution, equilibrium when early arrivals are forbidden, arrivals are allowed only at two points, and arrivals are allowed at three points. As can be seen in Table 8, the arrival probability at the opening time in equilibrium is much higher than the desired one. This phenomenon is stronger when the system is overloaded, with high λ and low μ, and in such cases the gap is narrowed by restricting arrivals the arrivals to two or three points. Table 9 compares the resulting expected waiting times.

		Continuous		2 points solution			
μ	λ	w	p_0	w	p_0	p_T	
8	10	.397	.635	.346	.553	.447	✓
10	10	.231	.470	.259	.518	.482	
12	10	.148	.355	.211	.507	.493	
14	10	.100	.280	.179	.502	.498	
15	10	.083	.249	.167	.501	.499	
16	10	.068	.218	.156	.501	.499	
18	10	.050	.180	.139	.501	.499	
20	10	.039	.156	.125	.501	.499	
30	10	.012	.072	.083	.500	.500	
8	15	.895	.955	.745	.794	.206	✓
10	15	.555	.740	.429	.572	.428	✓
12	15	.355	.568	.330	.527	.473	✓
14	15	.238	.439	.274	.511	.489	
15	15	.198	.396	.254	.508	.492	
16	15	.166	.354	.237	.505	.495	
18	15	.118	.283	.209	.502	.498	
20	15	.088	.235	.188	.501	.499	
30	15	.027	.108	.125	.500	.500	
8	20	1.25	1	1.25	1	-	
10	20	1	1	1	1	-	
12	20	.690	.835	.495	.595	.405	✓
14	20	.478	.669	.383	.537	.463	✓
15	20	.399	.605	.349	.524	.476	✓
16	20	.331	.530	.322	.516	.484	✓
18	20	.238	,428	.282	.507	.493	
20	20	.170	.340	.252	.503	.497	
30	20	.049	.147	.167	.501	.499	

Table 6: Equilibrium without early arrivals vs. two-points equilibrium. The right-most column indicates whether the two-points solution reduces w.

		Continuous		3 points solution					
μ	λ	w	p_0	w	p_0	t_m	p_T	w_{apx}	
8	10	.397	.635	.319	.50	.50	.34	.238	✓
10	10	.231	.470	.210	.40	.50	.32	.154	✓
12	10	.148	.355	.162	.39	.44	.35	.105	
14	10	.100	.280	.133	.37	.63	.28	.075	
15	10	.083	.249	.124	.37	.67	.27	.064	
16	10	.068	.218	.109	.35	.45	.34	.056	
18	10	.050	.180	.095	.34	.52	.33	.042	
20	10	.039	.156	.091	.34	.74	.28	.033	
30	10	.012	.072	.057	.34	.76	.32	.012	
8	15	.895	.955	.745	.79	-	.21	.443	✓
10	15	.555	.740	.429	.57	-	.43	.294	✓
12	15	.355	.568	.293	.47	.50	.34	.205	✓
14	15	.238	.439	.217	.40	.51	.32	.147	✓
15	15	.198	.396	.195	.39	.53	.31	.126	✓
16	15	.166	.354	.183	.39	.62	.26	.109	
18	15	.118	.283	.166	.38	.34	.39	.083	
20	15	.088	.235	.132	.35	.34	.31	.064	
30	15	.027	.108	.085	.34	.34	.34	.023	
8	20	1.25	1	1.25	1	-	-	.681	
10	20	1	1	1	1	-	-	.466	
12	20	.690	.835	.495	.59	-	.41	.332	✓
14	20	.478	.669	.357	.50	.53	.34	.244	✓
15	20	.399	.605	.312	.47	.50	.35	.210	✓
16	20	.331	.530	.306	.49	.38	.45	.183	✓
18	20	.238	.428	.221	.40	.48	.34	.139	✓
20	20	.170	.340	.185	.37	.52	.32	.108	
30	20	.049	.147	.114	.34	.63	.32	.038	

Table 7: Equilibrium without early arrivals vs. three-points equilibrium. The right-most column indicates whether the three-points solution reduces w.

μ	$\lambda = 10$				$\lambda = 15$				$\lambda = 20$			
	p_{apx}	p_{eq}	p_{2p}	p_{3p}	p_{apx}	p_{eq}	p_{2p}	p_{3p}	p_{apx}	p_{eq}	p_{2p}	p_{3p}
8	.08	.63	.55	.50	.006	.95	.79	.79	.05	1	1	1
10	.08	.47	.52	.40	.006	.74	.57	.57	.05	1	1	1
12	.08	.35	.51	.39	.007	.57	.53	.47	.05	.83	.59	.59
14	.08	.28	.50	.37	.007	.44	.51	.40	.06	.67	.54	.50
15	.07	.25	.50	.37	.007	.40	.51	.39	.06	.60	.52	.47
16	.07	.22	.50	.35	.007	.35	.50	.39	.06	.53	.52	.49
18	.07	.18	.50	.34	.008	.28	.50	.38	.06	.43	.51	.40
20	.06	.16	.50	.34	.008	.23	.50	.35	.06	.34	.50	.37
30	.04	.07	.50	.34	.009	.11	.50	.34	.05	.15	.50	.34

Table 8: Probability of arrival at $t = 0$: p_{apx} - approximate optimal; p_{eq} - equilibrium without early arrivals; p_{2p} - arrivals at two points; p_{3p} - arrivals at three points

μ	$\lambda = 10$				$\lambda = 15$				$\lambda = 20$			
	w_{apx}	w_{eq}	w_{2p}	w_{3p}	w_{apx}	w_{eq}	w_{2p}	w_{3p}	w_{apx}	w_{eq}	w_{2p}	w_{3p}
8	.24	.40	.35	.32	.44	.89	.74	.74	.68	1.25	1.25	1.25
10	.15	.23	.26	.21	.30	.55	.43	.43	.47	1.00	1.00	1.00
12	.10	.15	.21	.16	.20	.35	.33	.29	.33	.69	.49	.49
14	.07	.10	.18	.13	.15	.24	.27	.22	.24	.48	.38	.36
15	.06	.08	.17	.12	.13	.20	.25	.19	.21	.40	.35	.31
16	.06	.07	.16	.11	.11	.17	.24	.18	.18	.33	.32	.31
18	.05	.05	.14	.09	.08	.12	.21	.17	.14	.24	.28	.22
20	.03	.04	.12	.09	.06	.09	.19	.13	.11	.17	.25	.18
30	.01	.01	.08	.06	.02	.03	.12	.08	.04	.05	.17	.11

Table 9: Expected waiting time: w_{apx} - approximate optimal; w_{eq} - equilibrium without early arrivals; w_{2p} - arrivals at two points; w_{3p} - arrivals at three points

7 Concluding remarks

This paper considers a non-stationary queueing model with opening and closing times. We characterize the equilibrium solution and describe the underlying equations that govern it. Since an analytical solution of this model is out of reach, we solve it numerically. Our model is a variation of the one solved by Glazer and Hassin (1983), and it is motivated by natural questions concerning the ability to reduce the waiting time of customers who act independently and aim to maximize their individual welfare. This line of research falls within the growing research on strategic behavior in queueing systems.

We show that excluding early arrivals doesn't yield a significant reduction in expected waiting time unless the system is very heavily loaded (i.e., it is open for a short time relative to the demand).

The optimal strategy can be approximated fairly well by uniform distribution in the open interval $(0, T)$ and positive probabilities p_0 and p_T, representing the probabilities for a customer to arrive at time 0 and T respectively. Such approximate solution can be characterized by two parameters only. We compare an approximate optimal solution with the equilibrium solutions. The ratio between the expected waiting time under equilibrium and under the optimal solution increases when the system becomes more heavily loaded.

Finally, we computed the expected waiting time in equilibrium when arrivals are restricted to time 0 or T, and when they are in addition allowed at one internal point. Improved results are obtained by these restrictions of the arrival instants when the system is heavily loaded. In these cases, when the system is open continuously in $[0, T]$, too many customers tend to arrive at $t = 0$ ignoring the effect their arrival has on those who arrive later. The two and three-points restrictions reduce the probability that a customer arrives at the opening instant and by this reduce the expected waiting time.

A nice feature of these results is the simplicity of their implementation, in contrast, for example, to the common way of controlling customers behavior through price mechanisms. In our case, such a mechanism could be a time-dependent entry fee that seems quite impractical. We leave the question of whether other practical alternatives exist for future research. Another question for future research is whether restricting arrival times to small sets can be useful in reducing customers' waiting time in other models, for example in the scheduled batch service considered by Glazer and Hassin (1987).

References

[1] E. Altman, B. Gaujal, and A. Hordijk, "Multimodularity, convexity, and optimization properties," *Mathematics of Operations Research* **25** 324-347

(2000).

[2] M. Arbatskaya, K. Mukhopadhaya, and E. Rasmusen, "The Parking Lot Problem," Indiana University, Kelley School of Business, Department of Business Economics and Public Policy (2007).

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

[4] N. J. Bearden, A. Rapoport, and D. A. Seale, "Entry times in queues with endogenous arrivals: Dynamics of play on the individual and aggregate levels," *Experimental Business Research* (A. Rapoport and R. Zwick Eds.) **55** 201-221 (2005).

[5] T. Cayirli and E. Veral, "Outpatient scheduling in health care: a review of literature," *Production and Operations Management* **12** 519-549 (2003).

[6] A. Glazer and R. Hassin, "?/M/1: on the equilibrium distribution of customer arrivals," *European Journal of Operational Research* **13** 146-153 (1983).

[7] A. Glazer and R. Hassin, "Equilibrium arrivals in queues with bulk service at scheduled times," *Transportation Science* **21** 273-278 (1987).

[8] P. Guo, J.J. Liu, and Y. Wang, "Intertemporal service pricing with strategic customers," *Operations Research Letters* **37** 420-424 (2009).

[9] B. Hajek, "Extremal splittings of point processes," *Mathematics of Operations Research* **10** 543-556 (1985).

[10] R. Hassin and M. Haviv, *To Queue Or Not To Queue: Equilibrium Behavior in Queueing Systems* Kluwer (2003).

[11] M. Hlynka, "Real Life Queueing Examples," http://www2.uwidsor.ca/ hlynka/qreal.html, (1985).

[12] G. Kaandorp and G. Koole, "Optimal outpatient appointment scheduling," *Health Care Management Science* **10**, 217-229 (2007).

[13] G. Koole and E. van der Sluis, "Optimal shift scheduling with global service level constraint," *IIE Transactions* **35** 1049-1055 (2003).

[14] B. Jansson, "Choosing a good appointment system—A study of queues of the type (D/M/1)," *Operations Research* **14** 292-312 (1966).

[15] A. Lago and C. F. Daganzo, "Spillovers, merging traffic and the morning commute," *Transportation Research Part B* **41** 670-683 (2007).

[16] M. A. Lariviere and J. A. van Mieghem, "Strategically seeking service: how competition can generate Poisson arrival," *Manufacturing & Service Operations Management* **6** 23-40 (2004).

[17] V. V. Mazalov and J. V. Chuiko, "Nash Equilibrium in the Optimal Arrival time problem," *Computational Technologies* **11** 60-71 (2006). In Russian.

[18] M. Ostrovsky and M. Schwarz, "Adoption of standards under uncertainty," *The Rand Journal of Economics* **36** 816-832 (2005).

[19] M. Ostrovsky and M. Schwarz, "Synchronization under uncertainty," *International Journal of Economic Theory* **2** 1-16 (2006).

[20] H. Ostubo and A. Rapoport, "Vickrey's model of traffic congestion discretized," *Transportation Research Part B: Methodological* **42** 873-889 (2008).

[21] 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 Behaviour & Organization* **55** 67-91 (2004).

[22] D. A. Seale, J. E. Parco, W. E. Stein, and A. Rapoport, "Joining a queue or staying out: Effect of information structure and service time on arrival and staying out decisions," *Experimental Economics* **8** 117-144 (2005).

[23] W.S. Vickrey, "Congestion theory and transport investment," *The American Economic Review* **59** 251-260 (1969).

[24] S. Wang and L. Zhu, "A dynamic queuing model," *The Chinese Journal of Economic Theory* **1** 14-35 (2004).

Refael Hassin is a professor of Operations Research at Tel Aviv University, specializing in Combinatorial Optimization and Queueing models involving strategic behavior of customers and servers.

Yana Kleiner is a software engineer at Intel Israel. She holds a B.Sc degree in Information Systems from the Technion - Israel Institute of Technology, and M.Sc. degree in Operations Research from Tel Aviv University.