Queueing Models and
Service Management
Vol. 1, No. 2, page 031-058, 2018

QMSM
© PU 2018

Social Cost of Deviation: New and Old Results on Optimal Customer Behavior in Queues

Moshe Haviv[1,*] and Binyamin Oz[2]

[1]Department of Statistics and the Federmann Center for the Study of Rationality
The Hebrew University of Jerusalem
Jerusalem 91905, Israel
[2]Department of Statistics
The University of Auckland
Auckland 1142, New Zealand

(Received December 2017; accepted March 2018)

Abstract: We revisit some of the classic optimization problems in single- and multi-server queueing systems. We look at these problems as strategic games, using the concept of social cost of deviation (SCoD), which is the extra cost associated with a customer who deviates from the socially prescribed strategy. In particular, we show that a necessary condition for a symmetric profile to be socially optimal is that any deviation from it, if done by a single customer, is suboptimal; that is, the corresponding SCoD is nonnegative. We exemplify this by characterizing the socially optimal strategies for unobservable and observable "to queue or not to queue" problems and for multi-server selection problems. We then use the SCoD concept to derive the symmetric socially optimal strategy in a two-person game of strategic timing of arrival. Furthermore, we show that this strategy is also the symmetric Nash equilibrium strategy if the service regime is of random order with preemption.

Keywords: Social cost of deviation, strategic behavior in queues, social optimization.

1. Introduction and Preliminaries

The first thing that comes to mind when thinking of queues is waiting. Indeed, most if not all of the literature on queues deals with this issue or with the closely related issue of queue lengths. Of course, when the expectations of these two are a matter of concern, then we know by Little's rule that they are in fact two sides of the same coin. Expected waiting time is the long-run average time customers spend in the system from arrival to departure. This time includes the mean service time but also queueing time which is a result of others keeping the server busy. Waiting can be looked at as a constant sum game: on average, the waiting of a tagged customer due to the presence of others coincides with the waiting of other customers due to the presence of the tagged customer. The latter is known as the externalities that the tagged customer inflicts on the other users of the system. On average, queueing and externalities coincide. However, when social optimization is considered, one

* Corresponding author
Email : haviv@mscc.huji.ac.il

should think about the margins, that is, the marginal social effect of the behavior of a customer in a given system. For example, the marginal effect of a customer who decides to join a queue consists of two parts: (1) the waiting time of the arriving customer and (2) the externalities in the form of the aggregate added waiting time of others due to her arrival. While the former marginal effect coincides with its long-run average, this is not the case with respect to the latter marginal effect. Adding one more customer to a congested situation usually leads to much greater externalities than those inflicted on average, which, as we said, coincide with the expected queueing time of the extra customer herself.

This observation is the essence of why selfish interests and social interests deviate from each other when the issue of whether or not to join a queue is considered. Specifically, it is usually assumed that selfish gain and social gain due to service completion coincide, but regarding the issue of loss due to waiting, the marginal effect on the aggregate social cost is usually greater than the individual cost. In particular, in the case where the gain due to service completion is in between these two costs, selfish customers will decide to join a queue, whereas social optimization leads to the opposite decision.

This paper has two objectives. The first is to re-examine decision making in queues through the lens of social cost of deviation (SCoD), defined as the added social cost due to the deviation of a single customer from the socially prescribed strategy. In particular, we show how social optimization and SCoD are closely related. We exemplify this relation mostly in the context of whether or not to join a queue. The second objective is to state and prove the following new result: the socially optimal symmetric strategy in a symmetric game is such that letting one player deviate from it does not lead to any improvement. This result, when stated in terms of SCoD, turns out to be very useful in characterizing the optimal symmetric strategy. We exemplify this for a problem when two customers decide when to arrive for service during a given time interval.

The rest of the paper is organized as follows. Section 1.1 contains some well-known formulas from the field of queueing theory. In Section 2 we define social cost of deviation (SCoD). In Section 3 we consider a standby customer, namely, a customer who receives service only when the server would have been idle in her absence, and compare her expected waiting time with the marginal waiting associated with an arbitrary arrival. Section 4 contains a review of the existing literature on regulation of unobservable and observable queues. Section 5 deals with the problem of socially optimal routing strategies in multi-server selection models and its relation to the SCoD concept. In Section 6 we show that in the case of a symmetric optimization environment where mixed strategies are allowed, a necessary condition for a symmetric strategy to be socially optimal is that an individual deviation from it must be suboptimal; that is, the SCoD associated with any strategy must be non-negative. In Section 7 we deal with a strategic timing of arrival model with two customers. We use the SCoD concept to derive the symmetric socially optimal strategy, and to show that under a random service regime the (selfish) equilibrium behavior is in fact socially optimal. Section 8 concludes.

1.1. Preliminaries

With the exception of Section 7, we deal throughout with the M/G/1 model: customers arrive at a single-server queue according to a Poisson process with rate λ. Service times follow some general distribution with first and second moments $\overline{x}$ and $\overline{x^2}$, respectively. For stability it is assumed that $\lambda\overline{x}$, a value denoted by ρ, is strictly less than one. This is also the expected number of arrivals per service and the proportion of time in which the server is busy. Hence, it is usually referred to as the server utilization level. In the case where service times follow an exponential distribution, the rate of service is denoted by μ. In particular, $\overline{x} = 1/\mu$ and $\overline{x^2} = 2/\mu^2$. The model in this case is referred to as M/M/1. The Khintchine–Pollazcek formula says that in the case of a first-come first-served (FCFS) service policy, the mean queueing time (service exclusive) equals

$$W_q^{\mathrm{FCFS}} = \frac{\lambda\overline{x^2}}{2(1-\rho)}, \tag{1}$$

which in the M/M/1 case turns out to be

$$\frac{\rho}{\mu(1-\rho)}, \tag{2}$$

and the total waiting time (service included) in that case equals, and henceforth is denoted by,

$$W^{\mathrm{FCFS}} = \frac{1}{\mu(1-\rho)}. \tag{3}$$

Note that all these formulas hold not only for the FCFS case but also for any work-conserving, non-anticipating, and without preemption service discipline.[1] In the case of last-come first-served with preemption (LCFS-PR) where an interrupted service, when resumed, goes on from the point of last interruption, or in the case of a processor-sharing (PS) service discipline, the waiting time equals, and henceforth is denoted by,

$$W^{\mathrm{PS}} = \frac{\overline{x}}{1-\rho}. \tag{4}$$

Also, by Little's rule, the mean number of customers in the corresponding queue or system can be computed by multiplying each of the above expressions by λ.

[1] In the case of an M/M/1 system, the assumption of non-preemption can be dispensed with.

2. Social Cost of Deviations and Externalities in Single Server Queues

Definition 2.1. *Consider a symmetric game and two strategies S_1 and S_2. The social cost of deviation (SCoD) from S_1 to S_2, denoted by* $\mathrm{SCoD}(S_2,S_1)$*, is the difference between the sum of all individual expected costs resulting from two scenarios: (1) one arbitrary player uses strategy S_2 while all the other players use S_1, and (2) all the players use strategy S_1.*

In the queueing context, described in detail in the following sections, decision making involves comparing the benefit from service with the loss from waiting. Therefore, a key parameter of a queueing system that needs to be computed to determine SCoD is the expected aggregate added waiting time that an arbitrary arrival imposes on society as a whole (her inclusive). This is the expected difference in the total waiting time between a system with this arrival and a simulated one without it. We refer to this parameter as *marginal waiting*. In Haviv and Ritov [19] there are expressions for marginal waiting under various models. For the M/G/1 model under FCFS and under LCFS-PR marginal waiting equals

$$\overline{x}+\frac{\lambda\overline{x^2}(2-\rho)}{2(1-\rho)^2} \tag{5}$$

and

$$\frac{\overline{x}}{(1-\rho)^2}, \tag{6}$$

respectively. These two expressions coincide in the M/M/1 case and turn out to equal

$$\frac{1}{\mu(1-\rho)^2}, \tag{7}$$

which also holds for the M/M/1 model under the PS service regime.

Note the denominator of the marginal waiting expressions above. It comes with an extra inflating scale of $1/(1-\rho)$ in comparison with the mean waiting time. For example, in the M/M/1 case where $\mu=1$ in a system whose utilization level equals 0.9, one needs to spend 10 time units in the system on average but adds 100 time units of waiting to society in total (for only one time unit of service!). The externalities that an extra customer inflicts on others here equal 90 and they are spread among a number of customers. In the FCFS case, the expectation of this number is 90 as well, and it equals the expected number of customers who arrive after the arriving customer but before the first server idleness.[2]

[2] This can be proved as follows: upon arrival, the expected number of customers in the system (including the arriving customer) is 10 and therefore the time till the first idleness is distributed as 10 independent busy periods. During each of these independent busy period an expected number of $1/(1-\rho)=9$ customers arrive. Each such arriving customer incurs an extra waiting time of the same length as the service time of the

3. Marginal Waiting and Standby Customers

The value of W^{FCFS} coincides with the mean amount of work in the system upon arrival of a new customer (inclusive of the just-added service time). Since this parameter is not a function of the queue regime (as long as it is work-conserving), it applies to all models we consider here. Consider now the mean time in the system of a standby customer defined as a tagged customer who is singled out to receive service only when the server would have been idle without her. The service of such a customer might be preempted, but it is resumed from the point where it was last interrupted. Clearly, the expected waiting time of a standby customer coincides with the expected time from her arrival until the server is idle for the first time. Note that this parameter is also invariant with respect to the queue regime (again, assuming work conservation) and it is well known to be equal to $W^{\text{FCFS}}/(1-\rho)$. See, e.g., Haviv [13], p. 63.

It is tempting to jump to the conclusion that the mean time in the system of a standby customer coincides with the marginal waiting time as such a customer does not inflict any externalities. This is indeed true when service times are exponential. However, it is not always the case as the former is invariant with the queue regime, while the latter, as we will show shortly, is not thus invariant when preemptions are allowed. What is the reason behind this difference? The answer is that the extra preemptions that are associated with a standby customer have an effect on the overall performance of the system in comparison with the case without preemptions. In principle, and in fact this is the case, the effect of preemptions can be negative or positive and, of course, a zero effect is not ruled out.

Thus the next parameter we are after is the difference between the marginal waiting time and the expected time in the system for a standby customer. First observe that in the case of an M/M/1 queue this difference equals zero, as expected: in the case of exponential, i.e., memoryless service, preemptions do not have any effect on the performance of the system as a whole. A natural question arises: is exponential service also a necessary condition for this equality? The answer, as we show next, is no. In the following two examples, where M/G/1 is under the FCFS or the LCFS-PR regimes, equality is achieved if and only if $\overline{x^2}=2\overline{x}^2$, which of course holds in the M/M/1 case.

FCFS: Recalling (5) and (1), the difference in the FCFS case equals

$$\overline{x}+\frac{\lambda\overline{x^2}(2-\rho)}{2(1-\rho)^2}-\left(\overline{x}+\frac{\lambda\overline{x^2}}{2(1-\rho)}\right)(1-\rho)^{-1}=\frac{\rho}{1-\rho}(\overline{x}-\overline{r}),$$

where $\overline{r}=\overline{x^2}/2\overline{x}$ is the mean residual service time. In particular, this difference equals zero if and only if $\overline{x^2}=2\overline{x}^2$.

extra customer, i.e., 1.

LCFS-PR: Recalling (6), the difference in the LCFS-PR case equals

$$\frac{\overline{x}}{(1-\rho)^2}-\left(\overline{x}+\frac{\lambda\overline{x^2}}{2(1-\rho)}\right)(1-\rho)^{-1}=\frac{\rho}{(1-\rho)^2}(\overline{x}-\overline{r}),$$

which again equals zero if and only if $\overline{x^2}=2\overline{x}^2$.

3.1. Conditional externalities in M/G/1 queues

Clearly, in the stochastic queueing environment different customers not only experience different waiting times but also inflict different externalities. In particular, some information, or a signal, on the queueing process leads to the corresponding conditional expected externalities. A natural question that arises is, what is the expected externalities given one's service requirement?

We next state four such values for the M/M/1 FCFS model: the first three appear in Haviv [15], while the fourth appears in Mendelson and Wang [22] and Haviv and Ritov [19]. Consider an arbitrary customer and let E, W, L_a, L_d, and S be the (realization of the) externalities she inflicts, her waiting time, the queue length upon her arrival, the queue length upon her departure, and her service time, respectively. Then the conditional expected externalities she inflicts are given by

$$\mathbb{E}(E\,|\,W)=\frac{\rho}{1-\rho}W$$

$$\mathbb{E}(E\,|\,L_a)=\frac{L_a}{\mu(1-\rho)}-\frac{L_a}{\mu}$$

$$\mathbb{E}(E\,|\,L_d)=\frac{L_d}{\mu(1-\rho)}$$

$$\mathbb{E}(E\,|\,S)=\frac{\lambda}{2(1-\rho)}S^2+\frac{\rho^2}{(1-\rho)^2}S.$$

It is interesting to observe that the first three functions are linear while the fourth is quadratic (with a zero free coefficient). It is noteworthy that the fourth result is generalized to the M/G/1 case. In particular, the expected conditional externalities given S equal (see Haviv and Ritov [19])

$$\frac{\lambda}{2(1-\rho)}S^2+\frac{\lambda^2\overline{x^2}}{2(1-\rho)^2}S.$$

4. Regulation of Single-server Queues

We next deal with the most basic and most important decision in a queue: to join or not to join. This decision problem goes back to Naor [23] and Edelson and Hildebrand [5].

See Hassin and Haviv [10] and Hassin [9] for a survey of the literature. The model assumes an M/M/1 queue where each service completion leads to an individual reward of R but customers suffer a cost of C per unit of time in the system (service inclusive). Customers decide whether or not to join. They do so by comparing R with their expected waiting cost in the case where they join. Note that with regard to the waiting cost, they need to take into account decisions that were made, or are going to be made, by others. Without loss of generality, we assume that not joining comes with no cost and no reward. There are two versions of this problem. One is the unobservable case, where customers do not inspect the queue when they make up their mind (see Edelson and Hildebrand [5]), and the other is the observable version, where they do (see Naor [23]). We begin with the former case.

4.1. Individual vs. social optimization: The unobservable case

In order to avoid trivialities, assume that $R > C / \mu$ (as otherwise no one joins, even when the system is empty) and that $R < C / (\mu(1-\rho))$ (as otherwise it is optimal to join even when all others join). Customers face a symmetric non-cooperative game with two pure strategies: to join and not to join. A Nash equilibrium here is a possibly mixed strategy that, if used by everyone, is a (not necessarily unique) best response for an individual. Our assumptions on the cost and reward parameters lead to the fact that both “everyone joining" and “no one joining" are not equilibrium profiles. Hence, we look for an equilibrium that is based on mixing. Denote the joining probability under a mixed strategy by p. The equilibrium mixed strategy, denoted by p_e, is a joining probability such that, if used by everyone, no one can do better by using some other strategy. In particular, an individual is indifferent between joining and not joining under this scenario. This leads to p_e being the (unique) solution in p for

$$R - \frac{C}{\mu(1-p\rho)} = 0.$$

This indicates that

$$\lambda p_e = \mu - \frac{C}{R}$$

is the equilibrium arrival rate (which is not a function of λ as long as the right-hand side is nonnegative, which is in fact assumed).

From the social point of view, the equilibrium strategy is quite poor: it leaves those who join, as well as those who do not join, with zero utility. Clearly, reducing p to any positive value below p_e leads to a positive consumer surplus. So the next question is what is the socially optimal joining probability, denoted and defined by

$$p_s = \arg\max_{0<p<p_e} \left\{ p\left(R - \frac{C}{\mu(1-p\rho)} \right) \right\}.$$

The first-order condition of this optimization problem is

$$R = \frac{C}{\mu(1-p_s\rho)^2}, \tag{8}$$

which can be interpreted as follows. From (7) we learn that under the socially optimal scenario, the social gain from joining, R, equals the marginal waiting cost; that is, society is indifferent to whether or not a marginal customer joins. Rearrangement of (8) indicates that

$$\lambda p_s = \mu - \sqrt{\frac{C\mu}{R}} \tag{9}$$

is the socially optimal arrival rate (which is not a function of λ as long as the right-hand side is nonnegative, a fact that easily follows from our assumption that $R > C/\mu$).

To summarize, while under p_e *individuals* are indifferent between joining and not joining, under p_s *society* is indifferent.

4.2. Regulation by charging an entry fee

Society wishes everyone to behave in accordance with p_s; however, the standard assumption is that when left to themselves, users behave in accordance with p_e. In particular, when p_s is used the individual utility from joining is strictly positive so that not joining with some positive probability is individually suboptimal. To deal with this "tragedy of the commons" situation one needs to look for a way of modifying the rules (sometimes called *mechanism design*) in order to incentivize customers to follow the joining probability of p_s rather than p_e.

Perhaps the first solution that comes to mind is to charge an admission fee: everyone who joins pays some amount of money, denoted by T. This makes joining less favorable than before since from the customers' point of view, service is now rewarded by $R-T$ rather than by R. The optimal amount of such a fee, denoted by T_s, implements p_s as the equilibrium joining strategy and hence is the (unique) solution in T of

$$R - T - \frac{C}{\mu(1-p_s\rho)} = 0,$$

which leads to

$$T_s = R - \sqrt{\frac{CR}{\mu}}.$$

However, and even more interesting, recalling (8) immediately leads to

$$T_s = \frac{C}{\mu(1-p_s\rho)^2} - \frac{C}{\mu(1-p_s\rho)};$$

that is, everyone who joins pays the externalities they inflict on society (the marginal waiting cost minus their own share of this cost) under the assumption that p_s is the actual joining probability. Such a fee, when everyone pays the externalities they cause, is known in the economic circles as a Piguvian tax Pigou [24].

An alternative scheme is to sign a binding contract with a joining customer that charges a payment $f(X)$, where X is some random variable (whose value is unknown to the customer upon signing) and where $f(\cdot)$ is some real function. In the case where $\mathbb{E}(f(X)) = T_s$, regulation is achieved. There is no limit with respect to selecting such schemes but those that are more appealing come with an X that has something to do with the queueing process experienced by the signing individual. This can be for example the signing individual's service time. Then, by choosing $f(X)$ to be the conditional expected externalities given X, $\mathbb{E}(f(X)) = T_s$ is guaranteed by the law of total expectation.[3]

Recalling the discussion in Section 3.1, we consider four options for X: W, L_a, L_d, and S, and the conditional externalities stated there. Since we are considering the socially optimal arrival rate, we need to use λp_s wherever λ appears in these formulas. In particular, from (9) we know that $\lambda p_s = \mu - \sqrt{C\mu / R}$. Further simplification leads to the following.

Theorem 4.1. *The following four contracts regulate the joining rate:*

1. $f(W) = C\left(\sqrt{R\mu / C} - 1\right)W$
2. $f(L_a) = C\left(\sqrt{R / C\mu} - 1/\mu\right)L_a$
3. $f(L_d) = \sqrt{CR / \mu}\, L_d$
4. $f(S) = \frac{\mu C}{2}\left(\sqrt{R\mu / C} - 1\right)S^2 + C\left(\sqrt{R\mu / C} - 1\right)^2 S$.

Two other regulating contracts are presented below. They are based on W and L_d but are not the expected conditional externalities. Nevertheless, it is easy to show that their expected values coincide with T_s.

Theorem 4.2. *The following two contracts regulate the joining rate:*

1. $f(W) = \frac{C\mu}{2}W^2 - CW$

[3] This is not the only option since, as mentioned, any $f(\cdot)$ with $\mathbb{E}(f(X)) = T_s$ will do.

2. $f(L_a) = \frac{C}{2\mu} L_a^2 + \frac{C}{2\mu} L_a .$

Proof. See Kelly [21].

An important feature of the contracts considered above is that potential customers decide whether or not to join and sign them *before* knowing their realized value. An alternative approach is a random fee whose realized value is revealed to the customers prior to making their joining decision. Such a random fee needs to deter the optimal fraction of customers from joining, as the following theorem states.

Theorem 4.3. *A random entry fee regulates the joining rate if and only if it is drawn from a distribution with CDF, denoted by* $F(x)$*, that satisfies*

$$F(T_s) = p_s .$$

Moreover, the resulting equilibrium behavior is to join if and only if the realized value of the fee is less than or equal to T_s .

Proof. See Haviv and Oz [17].

An advantage of the random fee is that unlike under the other charging schemes, customers are now left with some (distribution-dependent) strictly positive consumer surplus.

4.3. Regulating by auctioning for priority

Hassin [8] suggests the following regulating scheme for an unobservable queue. Upon arrival, customers can either join or not join. Of course, the latter option comes with no reward or penalty. In the case of joining, customers are given the option to pay as much as they wish and their priority in the queue will be based on their payments. Specifically, one who pays x has a preemptive priority over those who pay y, with $y < x$. Ties are broken randomly. In particular, no seniority in the queue is respected. Here too customers are engaged in a symmetric non-cooperative game, where a mixed strategy prescribes a joining probability and some distribution of payments over the nonnegative axis. The question then is what is the equilibrium mixed strategy. No matter how simple or complicated the resulting equilibrium is, from the social point of view there is only one thing that matters: the probability of joining prescribed by the mixed strategy. In Hassin [8] it is shown that this probability equals p_s and, in particular, that this mechanism leads selfish customers to behave in equilibrium in a socially optimal way.

Of secondary interest is the distribution of payments for priority. We state it next and claim that everyone who joins in fact pays the externalities she inflicts (given that everyone behaves in accordance with this profile). Specifically, the payment is a continuous random variable whose support is $[0,a]$ with

$$a = \frac{C}{\mu(1-p_s\rho)^2} - \frac{C}{\mu},$$

and whose cumulative distribution function $F(x)$ can be read from the condition

$$R - x - \frac{C}{(1-p_s\rho(1-F(x))^2} = 0,\ 0 \le x \le a.$$

We next argue for the assertion that the equilibrium joining probability coincides with p_s. First, observe that the equilibrium distribution cannot include atoms at any point, as otherwise one is better off paying infinitesimally more and gaining a quantum reduction in waiting by overtaking everyone who joins and pays the value of the atom. Also, zero is clearly in the support of the equilibrium payment continuous distribution: if the lower edge were strictly larger than 0, it would be better for an individual to deviate from it to zero as no priority would be lost and the payment would be strictly smaller. Consider now a customer who pays zero. She becomes a standby customer among all those who pay and join. Her utility equals $R - C/\mu(1-p\rho)^2$, where p is the proportion of those who join. But, in equilibrium, her utility (like everybody else's) ought to equal zero as this is the utility that comes with not joining that is a pure strategy in the support. Thus, recalling (8), p here equals p_s.

4.4. Regulating M/M/1 by selecting the service regime

The commonly used queueing regime is FCFS. Deviating from this regime and adopting another one may result in a different equilibrium that results in better (or worse) social welfare. As we have just seen, despite the fact that money transfers are involved, switching to a service policy based on auctioning priority leads to socially optimal behavior in the unobservable case. We next show that in both the unobservable and observable cases, there exist service regimes that result in social optimization and do not involve any money transfer.

4.4.1. The unobservable case

Haviv and Oz [17] propose the following regime, which we call preemptive random priority. Specifically, upon arrival a uniformly [0,1] random variable, denoted by U, is drawn. The arriving customer inspects this number and decides whether or not to join. In case she joins, the number drawn becomes her preemptive priority parameter (the lower the number, the higher the priority). We claim that the unique symmetric equilibrium strategy is to join if and only if $U \le p_s$. Specifically, it leads to the socially optimal arrival rate as the effective joining probability is $\mathbb{P}(U \le p_s) = p_s$. The reasoning is as follows. Assuming that everyone uses the above strategy, a joining customer with a drawn value of $U \le p_s$ is a standby customer with respect to other customers with lower values than hers, which are a fraction of U of the total arrival rate, and therefore her utility from joining equals

$$R-\frac{C}{\mu(1-U\rho)^2}$$

which, by (8), is nonnegative. Also, a joining customer whose $U \geq p_s$ becomes a standby customer with respect to the other joining customers and her utility is $R-C/(\mu(1-p_s\rho)^2)=0$, which is the same utility that comes from not joining. In particular, not joining is also her best response. A clear advantage of this scheme, on top of the absence of the money transfer property, is that the regulator does not need to know the model's parameters and, in particular, there is no need to adjust the scheme in case that their values change.

Some variations of this scheme are suggested in Haviv and Oz [17]. Clearly, informing the customers of any monotone increasing function of U, rather than U itself, won't matter: if the function is $g(\cdot)$, then the threshold value will now be $g(p_s)$, rather than p_s. An attractive option is to take $g(U)=1/(\mu(1-U\rho)^2)$. The rationale is that $g(U)$ is now the expected waiting time for a customer who holds priority parameter U and joins, given that everyone who has higher priority than her joins as well. Another option is to have $g(U)=1/(\mu(1-U\rho)^2)$ for $U \leq p_s$ and $g(U)=1/(\mu(1-p_s\rho)^2)$ for $U \geq p_s$. Now the function $g(U)$ yields the expected waiting time of a customer who joins, given that everyone behaves in accordance with the equilibrium joining policy (which is also the socially optimal one). These two variations indeed lead to socially optimal behavior, but now the regulator needs to know some, or all, of the four model's parameters in order to implement them.

In the case where service is non-customized, for example, the server cooks burgers for hungry people who line up, there is no need for preemption. Specifically, upon service completion, the one who receives the completed good is the one who has the highest priority among all those present at that instant. The sample path of the queue is the same as in the case with preemptions due to the memoryless service distribution.

4.4.2. The observable case

Naor [23] studies the observable version of the above-mentioned decision problem. In fact, this paper preceded Edelson and Hildebrand [5] and it is rightly considered the paper that spawned the literature on customers' strategic behavior in queues. In his model, customers inspect the queue upon arrival and based on what they see, decide whether or not to join. The equilibrium analysis of this decision problem is rather simple and in fact is achieved by a dominant strategy which is to join if and only if n, the number seen upon arrival, is less than or equal to n_e-1, where

$$n_e = \max_n \left\{ n \geq 1,\, C\frac{n}{\mu} \leq R \right\}.$$

The issue of selecting the socially optimal threshold is in fact a Markov decision problem, where, in each state, composed of the number in the system upon an arrival, the manager decides whether or not to admit the new customer. Naor shows that the value of the objective function is a unimodal function with a peak and that the socially optimal threshold for the number in the system, denoted by n_s, equals $\lfloor x \rfloor$, where x uniquely solves

$$\frac{x(1-\rho)-\rho(1-\rho^x)}{(1-\rho)^2} = \frac{R\mu}{C}. \qquad (10)$$

For an alternative proof see Hassin and Haviv [10], p. 27–29.

The above condition can also be presented in terms of SCoD.

Theorem 4.4. *Recall definition 2.1. The optimal threshold strategy n_s is the unique integer, $n \geq 1$, that satisfies*

$$\text{SCoD}(n-1,n) > 0$$

and

$$\text{SCoD}(n+1,n) > 0.$$

Proof. To calculate the corresponding SCoD values we use the following results. Consider an M/M/1/ $(n+1)$ queue with exactly n customers in the system. Denote by T_n the expected time until the queue length reaches 0 or $n+1$, and by P_n the probability that it reaches 0 before it reaches $n+1$. The values T_n and P_n are derived in Hassin and Haviv [10], p. 28, using the gambler's ruin formulae, and are given by

$$T_n = \frac{n-(n+1)\rho\dfrac{1-\rho^n}{1-\rho^{n+1}}}{\mu(1-\rho)} \qquad (11)$$

and

$$P_n = \frac{1-\rho}{1-\rho^{n+1}}. \qquad (12)$$

Consider a customer who follows strategy $n-1$ while all the other customers follow strategy n. The aggregate social cost is affected, compared to the case where she too follows strategy n, only if she finds $n-1$ customers upon arrival. Denote the probability of this event by π_{n-1}. In this case she does not join, and the queue length remains $n-1$, while if she joins, the queue length becomes n. We next refer to the queue length processes under the former and latter scenarios as the actual and simulated processes, respectively. From the arrival instant of the tagged customer the two processes differ by 1 until they coincide for the first time. That happens when either one of these two events happen: (1)

the actual process reaches n, or (2) the simulated process reaches 0. Therefore, this period equals the aggregate waiting reduction under the actual process compared to the simulated one. Clearly, the expectation of the period coincides with T_n. Now, if this period ends when the simulated process reaches 0, then a reward of R is lost from the social gain, whereas if it ends when the actual process reaches n, the reward lost is offset by the last joining customer's reward, which is gained under the actual process but would have been lost under the simulated one. The probability of the former event is clearly P_n. In summary, the difference between the social cost under the actual process and that under the simulated process is

$$\text{SCoD}(n-1,n)=\pi_{n-1}(P_nR-CT_n).$$

Similarly, consider a customer who follows strategy $n+1$ while all the other customers are following strategy n. The aggregate social cost is affected only if she finds n customers upon arrival. Denote the probability of this event by π_n. In this event she joins, and the queue length becomes $n+1$, whereas if she does not join, the queue length remains n. Again, we refer to the queue length processes under the former and latter scenarios as the actual and simulated processes, respectively. The two processes differ by 1 until they coincide for the first time. For that to happen there must be a service completion first, which takes $1/\mu$ on expectation. Then, one of these two events must happen: (1) the simulated process reaches n, or (2) the actual process reaches 0. Once again, the expectation of the period is T_n. Therefore, the aggregate added waiting under the actual process is $1/\mu+T_n$. Now, if the two processes coincide for the first time when the actual process reaches 0, then an additional reward of R is gained, whereas if it happens when the simulated process reaches n, the reward gained by the additional customer is offset by the last rejected customer's reward, which is lost under the actual process but would be gained under the simulated one. The probability of the former event is P_n and, in summary, the difference between the social cost under the actual process and that under the simulated process is

$$\text{SCoD}(n+1,n)=\pi_n(C(1/\mu+T_n)-P_nR).$$

Now that the expressions for $\text{SCoD}(n-1,n)$ and $\text{SCoD}(n+1,n)$ are in hand, using (11) and (12) along with some simplification shows that $\text{SCoD}(n-1,n)>0$ if and only if

$$\frac{n(1-\rho)-\rho(1-\rho^n)}{(1-\rho)^2}<\frac{R\mu}{C},$$

and $\text{SCoD}(n+1,n)>0$ if and only if

$$\frac{(n+1)(1-\rho)-\rho(1-\rho^{n+1})}{(1-\rho)^2}>\frac{R\mu}{C},$$

which, by the monotonicity of the expression $n(1-\rho)-\rho(1-\rho^n)$ in n (proved in Naor

[23]), leads to the same condition in (10).

It is also possible to show that $\mathrm{SCoD}(k,n_s) > 0$ for all $0 \le k < n_s - 1$. We omit further details as the analysis uses similar arguments to those used in the proof above.

Observe that $n_s \le n_e$. The rationale behind this is similar to the rational behind $p_s < p_e$ in the unobservable version of this model. In both versions, customers, if left to themselves, cause congestion that is greater than desired. Indeed, customers who inspect the queue length and take into account only their own utility may join in cases in which society, which takes into account also the negative externalities they impose on others (in terms of making others wait longer due to their presence), would prescribe otherwise. The imposition of a toll may regulate the system. In fact, any entry fee, T, such hat

$$n_s = \max_n \left\{ n \ge 1, \frac{Cn}{\mu} \le R - T \right\},$$

leads to regulation.

Hassin [7] suggests a regulating policy for an observable M/M/1 queue by changing the service regime to any regime under which an arrival is placed anywhere but at the rear of the queue. Note that in the case where the arrival meets only one customer in the system, who of course is in service, she preempts the latter and commences service immediately upon arrival. An example of such a regime is LCFS-PR. Customers will be happy to join and the decision they face now is in fact when to renege. An assumption made here is that customers continuously monitor the queue length ahead of them. Indeed, if the queue ahead of them is too long, one is better off leaving for good. It is easy to see that under the resulting equilibrium, if anyone reneges, it will be the one at the rear of the queue. From the social point of view only one thing matters: how many customers remain in the queue after someone reneges. If this number is always n_s, social optimality is achieved. Hassin argues that this is indeed the case. The reasoning is as follows. The customer at the rear generates no externalities on others since under this scheme she will always be at the back of the line. Hence, her considerations of costs and rewards coincide with those of society. As society prescribes no more than n_s in the system and will order her to leave (indeed, society does not care who will be the one to renege), she will reach the same conclusion. Note that the implementation of this scheme does not involve any money transfer and does not require the manager to know any of the model's parameters.

Haviv and Oz [16] suggest another scheme. Suppose that there are an infinite number of ordered waiting slots. An arrival inspects them and learns which of them are occupied and which are vacant. She has the option to leave for good (balk) or to select one of the empty slots. Once a customer selects a slot, she stays there until her service is completed. In particular, changing slots or reneging later are not allowed. The server always serves the customer who is at the lowest-indexed slot among those that are occupied. This priority is kept in a preemptive manner.[4] The first question to ask is, what is the equilibrium strategy? The answer is as follows: an arrival joins the lowest-indexed empty slot as long as its index is less than or equal to n_s. It is possible to see that the same logic that explains Hassin's LCFS-PR regime holds here as well. In particular, social optimality is achieved with this scheme that does not require any knowledge of the the system's parameters from the operator's side, nor is there a need for any money transfers. Another advantage we have here is that no service is granted to customers who may renege later. For three more schemes, one based on charging a queue-dependent entry fee, one based on charging for a priority level, and one based on concealing the queue length and charging a constant fee, see Chen and Frank [4], Aleperstein [1], and Hassin and Koshman [12], respectively.

5. Server-selection Problems

Consider a multi-server system with a common arrival process. Specifically, there exists a Poisson stream of arrivals at rate λ. Each arriving customer selects one out of n exponential servers, where server i serves at a rate of μ_i, $1 \le i \le n$. Server selection is done in an unobservable fashion and without later regrets. A symmetric strategy profile is thus a probability distribution $P = (p_1, \dots, p_n)$ over the servers, such that server i is selected with probability $p_i \ge 0$, $1 \le i \le n$, and $\Sigma_{i=1}^{n} p_i = 1$. For ease of exposition, pure strategies are next denoted by an integer such that under strategy j, $1 \le j \le n$, server j is selected with probability one. These selections are done independently across the customers and hence the arrival process to server i is Poisson at rate λp_i, $1 \le i \le n$. Moreover, these n processes are independent. We next deal with two decision models. The first is where each server forms an M/M/1 queue and the second is where each forms an M/M/1/1 loss system.

5.1. Single-server queues

The following model was introduced in Bell and Stidham [3]. See also Hassin and Haviv [10], p. 62–64. In this model each of the above-mentioned single-server systems operates as an M/M/1 queue, where $\mu_1 \ge \mu_2 \ge \cdots \ge \mu_n$ is assumed without loss of generality. The socially optimal routing strategy minimizes the total expected number of customers in the system; that is, the optimization problem is

$$\min_{p_i, 1 \le i \le n} \sum_{i=1}^{n} \frac{\lambda p_i}{\mu_i - \lambda p_i}$$

$$s.t. \qquad \sum_{i=1}^{n} p_i = 1$$

[4] In case of non-customized service, the one who receives the completed product is the one at the lowest-indexed occupied slot. Of course, preemption is not an issue here.

$$p_i \geq 0,\ 1 \leq i \leq n.$$

The first-order conditions of this constrained optimization problem are as follows. For some value $\theta > 0$ (which is the Lagrange multiplier of the equality constraint) and for each $1 \leq i \leq n$, if $1/\mu_i > \theta$ then $p_i = 0$, and

$$\frac{1}{\mu_i(1-\lambda p_i/\mu_i)^2} = \theta \tag{13}$$

otherwise. In particular, since μ_i are ordered, there exists some index i_s, $1 \leq i_s \leq n$, such that only servers i, $1 \leq i \leq i_s$, operate, namely, their optimal p_i are positive, while all the other servers (if any) come with a zero socially optimal p_i, $i > i_s$. Note that the former class is never empty, while the latter may be so. Indeed, this is the case when λ is large enough.

The first-order condition is consistent with the following intuitive explanation. Under the socially optimal routing strategy, a marginal arrival is routed with positive probability to servers to which her joining will incur the minimal marginal waiting. This is true since the marginal waiting time associated with joining server i is $1/\mu_i(1-\lambda p_i/\mu_i)^2$ if $p_i > 0$ (see (7)) and $1/\mu_i$ otherwise; as in the latter case, the marginal waiting time consists only of the service time of the joining customer herself. It can also be seen that given that all behave in accordance with the socially optimal routing strategy, the SCoD from it to any pure (and hence, any mixed) one is nonnegative.[5] We do not give further details here since they would be quite similar to those given in the next section, which also deals with routing, but to loss systems rather than to queues.

The first order condition in (13) coupled with $\sum_{i=1}^{i_s} p_i = 1$ leads to

$$\theta = \frac{(\sum_{i=1}^{i_s}\sqrt{\mu_i})^2}{(\sum_{i=1}^{i_s}\mu_i - \lambda)^2}$$

and then the socially optimal arrival rate to server i is

$$\lambda p_i = \mu_i - \frac{\sqrt{\mu_i}}{\sqrt{\theta}} = \mu_i - \frac{\mu_i}{\sum_{j=1}^{i_s}\sqrt{\mu_j}}(\sum_{j=1}^{i_s}\mu_j - \lambda), 1 \leq i \leq i_s.$$

Finally,

[5] In fact, it is zero in case of joining server i, $1 \leq i \leq i_s$, and strictly positive otherwise.

$$i_s = \min\left\{ j : \frac{1}{\mu_{j+1}} \geq \frac{(\sum_{i=1}^{j}\sqrt{\mu_i})^2}{(\sum_{i=1}^{j}\mu_i - \lambda)^2}, 1 \leq j \leq n \right\}$$

where $\mu_{n+1} \equiv 0$. It is possible to see that the resulting value for θ is the Lagrange multiplier of the constraint $\sum_{i=1}^{n} p_i = 1$ under the optimal solution. Indeed, its value is the shadow price of this constraint.

5.2. Loss systems

The following model and the results that are surveyed here appear in Anily and Haviv [2]. We now assume that each of the servers is an M/M/1/1 loss system. Hence, under steady-state conditions a customer who seeks service from server i receives it (immediately) with probability $\mu_i/(\lambda p_i + \mu_i)$, $1 \leq i \leq n$. The complementary probability is that she is lost and never receives service. We also assume that receiving service from server i is valued at α_i, $1 \leq i \leq n$, and, without loss of generality, $\alpha_1 \geq a_2 \geq \cdots \geq \alpha_n$. The social objective is to maximize the expected gain per unit of time (or per customer). Hence, it can be put as

$$\begin{aligned} &\max_{p_i, 1\leq i\leq n} \sum_{i=1}^{n} \lambda \frac{p_i \mu_i}{\mu_i + \lambda p_i} \alpha_i \\ &s.t. \quad \sum_{i=1}^{n} p_i = 1 \\ &\qquad p_i \geq 0, 1 \leq i \leq n. \end{aligned}$$

Here too, the first-order conditions have the following form. For some $\theta > 0$ and for all $1 \leq i \leq n$, if $\alpha_i < \theta$ then $p_i = 0$. Otherwise,

$$\frac{\mu_i^2}{(\mu_i + \lambda p_i)^2} \alpha_i = \theta. \tag{14}$$

In particular, since α_i are ordered, there exists some index i_s, $1 \leq i_s \leq n$, such that only servers i, $1 \leq i \leq i_s$, come with strictly positive routing probabilities. The interpretation in terms of SCoD is also possible here.

Theorem 5.1. *The optimal routing strategy* P *satisfies*

$$\text{SCoD}(j, P) \geq 0, 1 \leq j \leq n,$$

with equality if $p_j > 0$.

Proof. We make use of the following lemma.

Lemma 5.2. *Consider an M/M/1/1 loss system with an arrival rate of* λ *and a service rate of* μ. *Assume that the service reward equals* α. *Consider an arbitrary arrival. The expected difference in the social gain between a case where she joins and a case where she does not equals*

$$\frac{\mu^2}{(\lambda+\mu)^2}\alpha. \tag{15}$$

In particular, if $\lambda = 0$, *this value equals* α.

Proof. The only arrival and service pattern under which there is a difference in the social gain under the two scenarios is if the following two events occur: the arriving customer finds an empty server, and then no one arrives before her service completion. In that case the difference in the social gain is α. Also, the former event happens with probability $\mu/(p\lambda+\mu)$, and as it turns out this is also the probability of the latter event. For more on this derivation see Anily and Haviv [2].

Consider a customer who follows the pure strategy j, $1 \le j \le n$, when all other customers use strategy P. The difference in the social gain (negative cost) compared to what it would have been if she used P is non-zero only if under the latter scenario she had been routed to server i, $i \ne j$. This is a p_i probability event. In that case, by Lemma 5.2 the gain lost by not having been routed to server i minus the gain due to being routed to server j instead equals

$$\frac{\mu_i^2}{(p_i\lambda+\mu_i)^2}\alpha_i - \frac{\mu_j^2}{(p_j\lambda+\mu_j)^2}\alpha_j.$$

Therefore,

$$\begin{aligned}\mathrm{SCoD}(j,P) &= \sum_{i\ne j} p_i\left(\frac{\mu_i^2}{(p_i\lambda+\mu_i)^2}\alpha_i - \frac{\mu_j^2}{(p_j\lambda+\mu_j)^2}\alpha_j\right)\\ &= \sum_{i\ne j,\,p_i>0} p_i\left(\theta - \frac{\mu_j^2}{(p_j\lambda+\mu_j)^2}\alpha_j\right)\end{aligned}$$

and the theorem follows since under the socially optimal strategy $\mu_j^2\alpha_j/(p_j\lambda+\mu_j)^2 \le \theta$ and if $p_j > 0$ then $\mu_j^2\alpha_j/(p_j\lambda+\mu_j)^2 = \theta$.

It is now clear that the first-order conditions have the following interpretation: an arriving customer is routed with a strictly positive probability only to servers that lead to the minimal SCoD, which equals zero.

The above leads to a computation procedure for finding i_s as well as the individual routing probabilities to servers i, $1 \le i \le i_s$. Specifically, (14) coupled with $\Sigma_{i=1}^{i_s} p_i = 1$ leads to

$$\theta = \left(\frac{\sum_{i=1}^{i^s} \mu_i \sqrt{\alpha_i}}{\sum_{i=1}^{i^s} \mu_i + \lambda} \right)^2$$

and

$$\lambda p_i = \mu_i \left(\sqrt{\frac{\alpha_i}{\theta}} - 1 \right), 1 \le i \le i_s. \tag{16}$$

Also,

$$i_s = \min \left\{ j : \alpha_{j+1} < \left(\frac{\sum_{i=1}^{j} \mu_i \sqrt{\alpha_i}}{\sum_{i=1}^{j} \mu_i + \lambda} \right)^2 , 1 \le j \le n \right\},$$

where $\alpha_{n+1} \equiv 0$.

Finally, note that the problem simplifies considerably in the case where the rewards are server independent. In this case all servers are active. Moreover, the arrival rates to the individual servers are proportional to their service rates, as can be inferred from (16).

6. The Equal SCoD Optimization Property

According to the definition of a Nash equilibrium profile, each (possibly mixed) strategy used by an individual player has the following property: all pure strategies in its support lead to the same cost, which is less than or equal to all costs achieved by pure strategies outside its support.[6] Besides the theoretical interest of this property, it has also much value in the computational search for equilibrium profiles: the just-stated conditions lead to a set of equalities and a set of inequalities (sets which vary with the support of the strategy), whose solutions (when they exist) are the equilibria one is after.

As shown for each of the models we have dealt with so far in this paper, a similar property holds for the socially optimal strategy where individual costs are replaced by the SCoD. This property is formally stated in Corollary 6.2 below. For example, in the unobservable M/M/1 problem it is socially optimal to join with probability p_s, that is, a mixed strategy where both “join" and “do not join" are in its support. Indeed, (8) implies that the SCoDs associated with these two strategies are similar, and equal zero. In the observable case, Theorem 4.4 implies that the socially optimal strategy is the only threshold strategy which the SCoD from it to all other strategies is positive. In other words, society admits only those whose net social cost is negative. Finally, in the two routing problems,

[6] The former class is never empty, while the latter class might be so.

(13) and (14) imply that those servers selected with positive probability, namely, those that are in the support of the socially optimal strategy, have the same zero SCoD, whereas those that remain closed have higher SCoDs.

In all the examples described above the characterization of the socially optimal strategy in terms of SCoD suggested in this paper is equivalent to the corresponding characterization that appears in the literature and is based on classic optimization and first order conditions. Nevertheless, the two approaches are fundamentally different. Classic optimization and first order conditions are based on the impact of an infinitesimal deviation from the socially prescribed strategy by *all* the customers, whereas the SCoD approach analyses the impact of the deviation of a *single* customer. The purpose of this section is to extend this idea beyond those cases where the equal SCoD property is revealed by means of multivariate optimization analysis. We next show that this property generally holds in symmetric games, where one looks for the socially optimal symmetric strategies.

Consider an N-player symmetric game, $N \geq 2$. Let $\mathrm{SC}_{(n_1,\ldots,n_k)}(S_1,\ldots,S_k)$ be the expected social cost when $n_i \geq 1$ players use (the possibly mixed) strategy S_i, $1 \leq i \leq k$, and $\sum_{i=1}^{k} n_i = N$. When all players use the same strategy S, the resulting social cost is hence denoted by $\mathrm{SC}_{(N)}(S)$. As before, the social cost of deviation from S_1 to S_2 is denoted by $\mathrm{SCoD}(S_2,S_1)$ and defined as the difference in the social cost between the case where all use S_1 but an arbitrary player uses S_2, and the case where all use strategy S_1. In the above notation,

$$\mathrm{SCoD}(S_2,S_1) = \mathrm{SC}_{(1,N-1)}(S_2,S_1) - \mathrm{SC}_{(N)}(S_1).$$

A socially optimal symmetric strategy is defined as a strategy S^* such that

$$S^* \in \arg\min_S \mathrm{SC}_{(N)}(S).$$

Suppose now that all players use S^* and the social planner is able to change the strategy used by one player. One might think that this added option leads to further improvement in the social cost due to the fact that the optimality of S^* is only among symmetric strategies. The following theorem shows otherwise; that is, under this scenario it is also optimal to assign S^* (or any pure strategy in its support) to the tagged player.

Theorem 6.1. *If S^* is an optimal symmetric strategy, then*

$$S^* \in \arg\min_S \mathrm{SCoD}(S,S^*). \tag{17}$$

Proof. Let S be some arbitrary strategy. Denote by S_ε the strategy under which one mixes between strategies S and S^*, giving probability ε to the former and probability $1-\varepsilon$ to the latter. Observe that the social cost when all use this mixed strategy is

$$\mathrm{SC}_{(N)}(S_\varepsilon) = \sum_{k=0}^{N} \varepsilon^k (1-\varepsilon)^{N-k} \frac{N!}{k!(N-k)!} \mathrm{SC}_{(k,N-k)}(S,S^*).$$

By the optimality of S^*, the minimum of this polynomial of degree N over $\varepsilon \in [0,1]$ is obtained at $\varepsilon = 0$. Since $\varepsilon = 0$ is a boundary solution, this implies that

$$0 \le \frac{d}{d\varepsilon} \mathrm{SC}_{(N)}(S_\varepsilon)\bigg|_{\varepsilon=0} = (\mathrm{SC}_{(1,N-1)}(S,S^*) - \mathrm{SC}_{(N)}(S^*))N = \mathrm{SCoD}(S,S^*)N$$

and the theorem follows since $\mathrm{SCoD}(S^*,S^*) = 0$.

Corollary 6.2. (The equal SCoD property) *A symmetric socially optimal strategy* S^* *satisfies*

$$\mathrm{SCoD}(s,S^*) = 0,\ s \in \mathrm{support}(S^*).$$

Proof. The condition in (17) implies that a necessary condition for S^* to be an optimal symmetric strategy is that for any other strategy, in particular, any $s \in \mathrm{support}(\mathrm{S}^*)$, $\mathrm{SCoD}(s,S^*) \ge \mathrm{SCoD}(S^*,S^*) = 0$. The fact that $\mathrm{SCoD}(S^*,S^*)$ is an expectation over the pure strategies in the support of the strategy in its first entry completes the proof.

The above remark can lead to a set of conditions that are obeyed (hopefully, uniquely) by S^*. We next exemplify this in the analysis of the socially optimal symmetric strategy when two customers choose strategically their arrival time to a queue.

7. Social Optimization and Regulation in Strategic Timing of Arrival

Another decision problem of interest is when to join the queue. This line of research was opened in Glazer and Hassin [6]. The assumption is that a (possibly random) number of customers seek service along some bounded time interval, say $[0,T]$ for some $T > 0$. Service continues after time T until the system is empty. Service times are random and for simplicity of analysis are assumed to be exponentially distributed with rate μ. As in all models dealt with here, customers suffer a waiting cost that is linear with the time they have to wait.[7] Assuming a FCFS regime, customers decide when to arrive so that their waiting will be as short as possible. There are usually two versions of this problem. One is where the seniority of those arriving before opening at time zero is respected, and the other where this is not the case (and all "early birds" arriving before opening enter service in random order). Since from the social point of view the option of being an early bird will never be exercised, we will assume that this option is not allowed.

Clearly, customers are involved in a non-cooperative game and one looks for a

[7] Later papers, e.g., [20] and [14], also added the feature of tardiness costs, reflecting how late service is rewarded in comparison with service that is granted at some ideal time, usually at $t = 0$.

symmetric equilibrium strategy that prescribes (a possibly random) time of arrival. This problem is solved in Hassin and Kliener [11] for the case of a Poisson distributed number of arrivals. The next question is what is the socially optimal symmetric arrival strategy. As it turns out, this is a harder question and only numerical results are reported in Hassin and Kliener [11]. Note, however, that under this criterion the regime used is irrelevant (as long as it is work-conserving and non-anticipating).

7.1. Social optimization in the case of two customers

The following theorem shows that the symmetric socially optimal strategy in the case of two customers prescribes them to arrive with positive probabilities at time 0 and time T, and with uniform density at any point in $(0,T)$. This result, which we analytically prove using the equal SCoD property, resembles the numerical results for the case of Poisson distributed number of customers reported in Hassin and Kliener [11].

Theorem 7.1. *The unique symmetric socially optimal strategy in the two-customer strategic timing of arrival model is as follows. There are two atoms of size* $1/(\mu T+2)$ *each at* 0 *and* T*, and there exists a uniform density of* $\mu/(\mu T+2)$ *along* $(0,T)$.

Proof. We begin the proof by stating and proving the following.

Lemma 7.2. *Consider a strategy* S *that has atoms at the points* $0, t_1, t_2, \ldots, T$. $0 < t_1 < t_2 < \ldots < T$, *of sizes (possibly zero)* $p_0, p_{t_1}, p_{t_2}, \ldots, p_T$, *respectively, and some density elsewhere (again, possibly of size zero) of* $f(x)$, $x \in [0,T]$. *Of course,* $p_0 + p_T + \sum_i p_{t_i} + \int_{x=0}^{T} f(x)dx = 1$. *Then, the social cost when one customer uses a pure strategy* t, *that prescribes arrival at time* $t \in [0,T]$ *with probability one, while the other customer uses strategy* S *equals*

$$\mathrm{SC}_{(1,1)}(t,S) = \frac{2}{\mu} + \frac{1}{\mu}\mathcal{P}_S(t) \tag{18}$$

where

$$\mathcal{P}_S(t) = e^{-\mu t}p_0 + e^{-\mu(T-t)}p_T + \sum_i e^{-\mu|t-t_i|}p_{t_i} + \int_{x=0}^{T} e^{-\mu|t-x|}f(x)dx. \tag{19}$$

Proof. The first term of (18), $2/\mu$, is the sum of the two expected service times. This is a cost that the two customers incur regardless of the strategies used by them or the process progression. Additional cost in the form of queueing time is incurred if and only if the two customers find themselves in the system at the same time. Conditioned on this event, due to the memoryless property of the exponential distribution and regardless of the service regime, an additional queueing time of exponential length with mean $1/\mu$ is incurred. It is therefore left to show that $\mathcal{P}_S(t)$ is the probability of the event that the two customers find themselves in the system at the same time. Conditioned on the event that the customer who uses strategy S arrives at time $x \in [0,T]$, the probability we are after is the probability that the service time of the customer who arrived first is greater than $|t-x|$, which equals $e^{|t-x|}$. Finally, integrating with respect to the distribution prescribed by strategy S completes the proof.

From Lemma 7.2 (see (18)) we learn that $\mathrm{SCoD}(t,S) = \mathrm{SC}_{(1,1)}(t,S) - \mathrm{SC}_{(2)}(S)$ is a function of t only through $\mathcal{P}_S(t)$. Hence, the optimality condition in Corollary 6.2 is equivalent to that an optimal strategy S^* is such that

$$s \in \arg\min_{t\in[0,T]} \mathcal{P}_{S^*}(t), \; s \in \mathrm{support}(S^*). \tag{20}$$

Straightforward differentiation of (19) shows that for any strategy S

$$\frac{d^2}{dt^2}\mathcal{P}_S(t) = \mu^2\mathcal{P}_S(t) - 2\mu f(t), \; t \in [0,T]. \tag{21}$$

A conclusion from (21) is that if S is an optimal symmetric strategy, then $f(t) > 0$ for all $t \in (0,T)$ since otherwise, namely, if $f(t) = 0$ along some interval, a point x in the interior of that interval is such that $x \notin \mathrm{support}(S)$ but comes with a lower value of $\mathcal{P}_S$, contradicting (20). This is the case since along such an interval, $\mathcal{P}_S(t)$ is strictly convex as $d^2\mathcal{P}_S(t)/dt^2 = \mu^2\mathcal{P}_S(t) > 0$. As a consequence, the support of the equilibrium strategy contains all points in $[0,T]$, which means that $\mathcal{P}_S(t)$ is constant and, in particular, $d^2\mathcal{P}_S(t)/dt^2 = 0$ for all $t \in [0,T]$. Equation (21) then implies that $f(t)$ is constant as well. Moreover, $\mathcal{P}_S(t)$, being constant, must be continuously differentiable, which excludes the possibility of atoms in the interior of $(0,T)$, i.e., $p_{t_1} = p_{t_2} = \ldots = 0$. The condition $\mathcal{P}_S(0) = \mathcal{P}_S(T)$ implies that $p_0 = p_T = p$ for some $0 \le p < 1/2$ and $f(t) = (1-2p)/T$. Finally, (21) evaluated at $t = 0$ along with $d^2\mathcal{P}_S(t)/dt^2 = 0$ imply that

$$\frac{\mu}{2}\mathcal{P}_S(0) = f(0). \tag{22}$$

This means that p uniquely solves

$$\frac{\mu}{2}\left[p + e^{-\mu T}p + \int_{x=0}^{T} e^{-\mu x}\frac{1-2p}{T}dx\right] = \frac{1-2p}{T},$$

which implies that the strategy S such that $p_0 = p_T = 1/(\mu T + 2)$ and $f(t) = \mu/(\mu T + 2)$ is the unique symmetric socially optimal strategy.

Remark 7.3. *The expected waiting time under the socially optimal strategy equals* $\mathrm{SC}_{(2)}(S^*)/2$ *and since* $0 \in \mathrm{support}(S^*)$, *it also equals*

$$\begin{aligned}\mathrm{SC}_{(1,1)}(0,S^*)/2 &= \frac{1}{\mu} + \frac{\mathcal{P}_{S^*}(0)}{2\mu} \\ &= \frac{1}{\mu} + \frac{f(0)}{\mu^2} = \frac{1}{\mu} + \frac{1}{\mu^2 T + 2\mu},\end{aligned}$$

where for the penultimate equality we use (22). This value is of course strictly less than the *corresponding equilibrium waiting time under FCFS, which is calculated in Haviv and Ravner* [18], *and it equals* $1/\mu + 1/(\mu^2 T + \mu)$.

7.2. Regulation in the case of two customers

As expected, in this decision model with externalities that vary with the arrival time, the equilibrium and the socially optimal strategies do not coincide. But there is an exception to this rule, as we show in the next theorem. This is the case where the number of customers equals two and the service regime is PS (or random with preemption (ROP)[8]). In particular, under these two regimes the system is self regulated.

Theorem 7.4. *When two customers decide when to arrive in the interval* $[0,T]$ *the symmetric socially optimal strategy coincides with the symmetric equilibrium strategy under the PS and the ROP regimes.*

Proof. A direct approach will be to compute the equilibrium strategy under PS and ROP and show that it is the same strategy described in Theorem 7.1. We would like to provide a more qualitative proof that uses the following observation. This observation in fact holds for any number of customers and hence has an interest of its own. It is stated and proved accordingly.

Observation 7.5. *Consider a time period between two departure or arrival events that corresponds to a time period where the number of customers in the system is constant. Fix that number at* k *and assume that* $k \geq 2$. *Under the PS regime the queueing time of any customer among these* k *customers, and during this period, equals the externalities she imposes on the other* $k-1$ *customers.*[9] *Likewise, under the ROP regime, the expected queueing time of any customer equals the expected externalities she imposes on others.*

Proof 1. Assume that the period under consideration is of length ℓ and tag one customer among the k present. The amount of service she gets under PS during this period is ℓ / k and the rest of the time, $\ell(k-1)/k$, is therefore her queueing time. Moreover, in the fraction of time in which this customer is in service, all other $k-1$ customers wait, causing an additional aggregate queueing time of $\ell(k-1)/k$. Likewise, under ROP, in the last event (departure or arrival) prior to this period a lottery is performed to decide who will be the one to enter service. The probability that the tagged customer will not be this one, and hence will have to wait during this period, is $(k-1)/k$ and hence her expected queueing

[8] Under this scheme random lotteries for who receives service next are performed only upon arrival and departure instants.

[9] In the case of a PS model, queueing is referred to as the slowdown due to having to share the server with others. For example, if three customers share the server for three mins, it is as if each one of them were queueing for two mins and receiving one minute of service.

time is $\ell(k-1)/k$. With the complementary probability, $1/k$, she will be the one to receive service and all $k-1$ others will wait for ℓ time units, causing an expected aggregate queueing time of $\ell(k-1)/k$.

Proof 2. Consider two customers, A and B, who are together for some time interval in a PS system. During this time interval, A has been in service exactly for the same length of time as B, and this is regardless of how many additional customers are present. Likewise, in the case of a ROP regime: A may affect B only if he wins the lottery deciding who enters next before B does, an 0.5 probability event, and this is regardless of how many earliest lotteries they both have lost to others. The same extra queueing time will be inflicted on A if B wins this lottery first.

It is tempting to deduce from Observation 7.5 that the overall queueing time of a customer in this model equals the overall externalities she imposes on others. However, this deduction is not true due to the fact that a customer might cause additional waiting of others during time periods after her departure. This is the case when her arrival time affects the queue length also after her departure, which in turn affects the waiting time of future arrivals. Note however that this scenario is possible only if the number of customers is greater than two. Hence, in the two-customer case we have the following:

Corollary 7.6. *Assume that the number of customers equals two and let $Q(S_1,S_2)$ and $E(S_1,S_2)$ be the expected queueing and expected externalities associated with a customer who uses strategy S_1 when the other customer uses strategy S_2. Then, for any strategy profile (S_1,S_2),*

$$Q(S_1,S_2)=E(S_1,S_2).$$

Hence, the resulting individual cost under this profile is

$$1/\mu+Q(S_1,S_2)=\frac{2/\mu+Q(S_1,S_2)+E(S_1,S_2)}{2}=\frac{\mathrm{SC}_{(1,1)}(S_1,S_2)}{2}. \tag{23}$$

We now conclude the proof of Theorem 7.4. The symmetric equilibrium strategy is S_e such that if used by one customer, it is a best response for the other customer. Specifically,

$$S_e\in\arg\min_S\{1/\mu+Q(S,S_e)\}.$$

On the other hand, Corollary 7.6 (see (23)) implies that this condition is equivalent to

$$S_e\in\arg\min_S\{\frac{\mathrm{SC}_{(1,1)}(S,S_e)}{2}\}=\arg\min_S \mathrm{SCoD}(S,S_e),$$

that, in this case, uniquely characterizes the symmetric socially optimal strategy.

8. Concluding Remarks

The Nash equilibrium solution concept provides a systematic way of dealing with non-cooperative games. In particular, when dealing with symmetric games, we look for a strategy such that when it is used by all, it is also one's best response. Despite its seemingly simple definition, characterization of the equilibrium strategy requires a thorough examination of individual incentives and rewards. Indeed, most if not all of the literature on strategic behavior in queues analyzes such symmetric games in the queueing context through the lens of the Nash equilibrium concept. Many of those studies also deal with the socially optimal counterpart, that is, a symmetric strategy profile that minimizes the aggregate cost among all players. This problem is typically approached in a straightforward way: write the social cost as a function of the symmetric strategy and optimize over the strategy space. In this paper we introduce the social cost of deviation method, which, unlike the traditional methods, preserves the incentives and rewards nature of each game. We show that the characterization of the socially optimal symmetric strategy is similar to that of the Nash equilibrium strategy, but instead of looking at the incentives of the individual we now look at those of society. In particular, the socially optimal strategy is such that when it is used by all, the social cost of deviation from it to any strategy is nonnegative. This characterization provides a unified approach to social optimization in symmetric games. Moreover, as exemplified in the proof of Theorem 7.1, the social cost of deviation may lead to a computational method for the socially optimal strategy in problems that otherwise require infinite-dimensional optimization (e.g., when the strategy space is all the distributions over a closed interval).

Acknowledgement

We thank the anonymous referee for her/his valuable comments that greatly improved this paper. This research was partly supported by Israel Science Foundation grant no. 511/15.

References

[1] Aleperstein, H. (1988). Optimal pricing for the service facility offering a set of priority prices. *Management Science*, 34, 666–671.

[2] Anily, S., & Haviv, M. (2017). The price of anarchy in a loss system. (work in progress).

[3] Bell, C. E., & Stidham, Jr. S. (1983). Individual versus social optimization in allocation of customers to alternativ servers. *Management Science*, 29, 831–839.

[4] Chen, H., & Frank, M. (2001). State dependent pricing with a queue. *IIE Transactions*, 33, 847–860.

[5] Edelson, N. M., & Hildebrand, D. K. (1975). Congestion tolls for Poisson queueing processes. *Econometrica*, 43, 81–92.

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

[7] Hassin, R. (1985). On the optimality of first come last served queues. *Econometrica*, 53, 201–202.

[8] Hassin, R. (1995). Decentralized regulation of a queue. *Management Science*, 41, 163–173.

[9] Hassin, R. (2016). *Rational Queueing*. CRS Press.

[10] Hassin, R., & Haviv, M. (2003). *To Queue or Not to Queue: Equilibrium behaviour in Queueing System*. Kluwer Kluwer Academic Publishers, Boston / Dordrecht / London.

[11] Hassin, R., & Kliener, Y. (2010). Equilibrium and optimal arrival patters to a server with opening and closing times. *IIE Transactions*, 43, 164–175.

[12] Hassin, R., & Koshman, A. (2017). Profit maximization in the M/M/1 queue. *Operations Research Letters*, 45, 436–441.

[13] Haviv, M. (2013). *Queues – A Course in Queueing Theory*, Springer.

[14] Haviv, M. (2013). When to arrive to a queue with tardiness costs. *Performance Evaluation*, 70, 387–399.

[15] Haviv, M. (2014). Regulating an M/G/1 when customers know their demand. *Performance Evaluation*, **77**, 57–71.

[16] Haviv, M., & Oz, B. (2016). Regulating an observable M/M/1 queue. *Operations Research Letters*, 44, 196–198.

[17] Haviv, M., & Oz, B. (2018). Self-regulation of an unobservable queue. *Management Science*, 64, 2380-2389.

[18] Haviv, M., & Ravner, L. (2015). Strategic timing of arrivals to a finite queue multi-server loss system. *Queueing Systems: Theory and Applications*, 81, 71–96.

[19] Haviv, M., & Ritov, Y. (1998). Externalities, tangible externalities and queue disciplines. *Management Science*, 44, 850–858.

[20] Jain, R., Juneja, S., & Shimkin, N. (2013). The concert queueing game: Strategic arrivals with waiting and tardiness costs. *Queueing Systems: Theory and Applications*, 74, 369–402

[21] Kelly, F. K. (1991). Network routing. *Philosophy Transactions of the Royal Society*, A337, 343–367.

[22] Mendelson, H., & Whang, S. (1990). Optimal incentive-compatible priority pricing for the M/M/1 queue. *Operations Research*, 38, 870–883.

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

[24] Pigou, A. C. (1920). *The Economics of Welfare*. Macmillan.