MANAGEMENT SCIENCE
Vol. 00, No. 0, Xxxxx 0000, pp. 000–000
ISSN 0025-1909 | EISSN 1526-5501 | 00 | 0000 | 0001

INFORMS
DOI 10.1287/xxxx.0000.0000
© 0000 INFORMS

Self-Regulation of an Unobservable Queue

Moshe Haviv, Binyamin Oz
Department of Statistics and Federmann Center for the Study of Rationality, The Hebrew University of Jerusalem, 91905
Jerusalem, Israel moshe.haviv@gmail.com binyamin.oz@gmail.com

We consider an unobservable M/M/1 queue where customers are homogeneous with respect to service valuation and cost per unit time of waiting. It is well known that left to themselves, in equilibrium, customers join the queue at a rate higher than is socially optimal. Hence, regulation schemes, under which the resulting equilibrium joining rate coincides with the socially optimal one, should be considered. We suggest a classification of regulation schemes based on a few desired properties and use it to classify schemes from the existing literature. To the best of our knowledge, none of the existing schemes possesses all the properties, and in this paper we suggest such a scheme. Its novelty is in assigning random priorities to customers, prior to their decision whether to join or balk. We also introduce variations of this regulation scheme as well as others that are also based on randomization.

Key words: Preemptive Random Priority; Regulation of a Queue; Strategic Behavior in Queue

1. Introduction

It is well known that left to themselves, customers, users, or commuters tend to overcrowd the facility or system they use. For example, it is possible that two-thirds of the potential users of a road decide to take it, while it is socially optimal that only one-third of them do so. The reason for that is that users usually mind only their own utility while ignoring the side effects, known as externalities, associated with the inconvenience they impose on others due to their decisions. Specifically, joining a crowded system is usually associated with increased holding or waiting time for others. Forcefully stopping users from joining is often perceived as unfair or unacceptable, and the question is therefore how to induce them to behave in a socially optimal way while still minding only their self-interests. Moreover, the less that is done or required in order to achieve that, the better. In the context of queues, there are a few ways to achieve this goal. For example, imposing an entry fee may deter some potential users from joining. In this way the system is regulated to function in a socially better way.

The principle of using entry fees to reduce queues was recognized by Leeman (1964). The first mathematical analysis in this direction was introduced in the pioneering paper of Naor (1969), where a formulation and analysis of equilibrium and welfare-maximizing joining strategies are derived for the observable cost-reward M/M/1 queueing model. In addition, the calculation of the entry fee that induces socially optimal behavior is given. The regulation of observable queues using queue-length-dependent fees is suggested in Alperstein (1988) and Chen and Frank (2001). Hassin (1985) suggests the use of the First-Come Last-Served service regime to regulate the system. This method is discussed in detail in Section 5.4 below. A major advantage of this scheme is that its implementation does not require money transfers or knowledge of the model parameters. For further discussion of these properties and some other schemes based on modified service regimes, see Haviv and Oz (2016b).

The unobservable version of Naor's model was introduced in Edelson and Hilderbrand (1975) where the socially optimal (and profit-maximizing) entry toll is derived. Regulation of unobservable M/M/1 queues is also studied in Haviv (2014) using entry contracts, and in Hassin (1995) using auctioning priority. These papers are discussed in some detail in Section 3.1 below. For a comprehensive overview of the literature on strategic behavior in queues, see Hassin and Haviv (2003). The main purpose of this paper is to contribute to this stream by suggesting new regulation schemes for the unobservable M/M/1 queuing model that are based on giving customers random priorities. These schemes can be viewed as the unobservable counterpart of Hassin (1985), as they are also robust to the model parameters and do not involve money transfers.

2. Model and Preliminaries

We deal here with the following decision model, which first appeared in Edelson and Hilderbrand (1975). The basic model is the M/M/1 queue. Customers seek service from a single server in front of which a waiting line is formed. Service times are exponentially distributed with a mean value of μ^{-1}. The arrival process is Poisson with potential rate of λ customers per unit time. The server utilization level equals λ/μ and is denoted by ρ. Assume that $\rho < 1$, i.e.,[1] $\lambda < \mu$. Each customer values service by R and incurs a waiting cost of C per unit time in the system (service included). Customers are risk neutral and interested in maximizing their expected monetary utility. Recall that $1/(\mu - \lambda)$ is the expected time in the system under any work-conserving and non-anticipating queue regime such as First-Come First-Served (FCFS), assuming all customers join.[2]

[1] It will later be shown that this condition is not required for equilibrium and social optimization analysis. It is imposed here for ease of exposition.

[2] By "non-anticipating" we mean that the choice of who gets service is not determined by the actual (past or residual) service times of the present customers. "Work conservation" means that the total amount of work in the system is the same as in a FCFS regime, everything else being equal. In particular, the server is never idle while customers are present. For more on these definitions see, e.g, Haviv (2013), p. 53.

Without loss of generality balking comes with a zero reward. In order to avoid trivialities we assume $R > C/\mu$ and $R < C/(\mu - \lambda)$. The first (respectively, second) assumption implies that if nobody (respectively, everyone) joins, then a selfish individual is better off joining (respectively, balking).

Suppose now that each customer decides whether to balk or join the queue (without any further information such as the queue length upon arrival or his service requirement). This decision-making problem is not an optimization problem but rather a non-cooperative game. Indeed, if all join (respectively, balk), one had better balk (respectively, join). Hence, mixed strategies must be considered and the symmetric Nash equilibrium is a possible solution concept to adopt. Specifically, we are looking for a joining probability such that if used by all then under the resulting steady-state conditions, this probability is also one's best response. Note that in the case where this probability is strictly between zero and one (as it is here), one is indifferent between joining or balking.

This implies that under Nash equilibrium all customers join with probability

$$p_e = \frac{\mu - \frac{C}{R}}{\lambda}, \tag{1}$$

as it solves uniquely the equation

$$R - \frac{C}{\mu - p\lambda} = 0$$

for p. Note that in equilibrium, those who join, as well as those who balk, end up with zero expected surplus. Zero is therefore the social welfare.

Note that $p_e\lambda < \mu$, whether or not $\lambda < \mu$. In fact, $p_e\lambda = \mu - C/R$; namely, the effective arrival rate in equilibrium is not a function of λ as long as $\lambda > \mu - C/R$.

Society, in the aggregate, would be better off if it (or a central planner on its behalf) controlled the joining probability. The socially optimal probability is

$$p_s = \arg\max_{0\le p\le 1} \left\{ p\lambda \left(R - \frac{C}{\mu - p\lambda} \right) \right\}. \tag{2}$$

In words, p_s is the joining probability that maximizes the mean social welfare gained per unit time. Of course, if for some reason the joining probability is larger (respectively, smaller) than p_s, the central planner would like fewer (respectively, more) customers to join. It is only under p_s that the central planner is indifferent whether an individual customer joins or not. This is certainly not the case under p_e: the social gain already equals zero and each additional customer who joins makes things even worse. The first-order condition of the central planner's problem (2) is

$$R - \frac{C}{\mu(1 - p\rho)^2} = 0 \tag{3}$$

and the solution is

$$p_s = \frac{\mu - \sqrt{\frac{C\mu}{R}}}{\lambda}. \tag{4}$$

Clearly,[3] $p_s < p_e$. Note that $p_s\lambda = \mu - \sqrt{C\mu/R}$; namely, the socially optimal effective arrival rate is not a function of λ as long as $\lambda \geq \mu - \sqrt{C\mu/R}$.

2.1. Externalities and Standby Customer

We denote by a *standby* customer one who receives service only when the server would otherwise be idle. In particular, a standby customer is preempted by those arriving while he is receiving service. When there are no other customers in the system, the standby customer's service is resumed from the point where it was interrupted. It is well known (see, e.g., Haviv 2013, p. 64) that the mean time in the system for a standby customer in a M/M/1 queue equals

$$\frac{1}{\mu(1-\rho)^2}, \tag{5}$$

and note that the mean time in the system for a standby customer when $p = p_s$ equals

$$\frac{1}{\mu(1-p_s\rho)^2} = \frac{R}{C}. \tag{6}$$

The concept of a standby customer is closely related to that of externalities. For a comprehensive discussion of this relation see Haviv and Oz (2016a). The externalities that a tagged customer imposes on the other customers are defined as the total added time that the other customers wait, compared to how long they would have waited if he had not joined. In order to compute this added time, we make this tagged customer a standby customer. By doing so we do not change the aggregate waiting time of all customers (when the tagged customer is considered as part of society), but now all the added waiting time is absorbed by this customer. Recall that the waiting time of a "normal customer," namely, $1/(\mu(1-\rho))$, is a cost that he would have incurred anyhow. The increase of his waiting time, which clearly equals

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

is defined as the externalities that the tagged customer imposes on the others.

Externalities are the root cause of the difference between the equilibrium behavior of customers and the socially optimal behavior. Consider for example the "to queue or not to queue" type of decision making we consider here. Although a joining customer inflicts (negative) externalities on

[3] This inequality can be derived not only by means of minimal algebra, but also from the fact that any $p > p_e$ leads to a negative social welfare and hence the socially minded p needs to be with $p < p_e$ in order to gain some positive surplus. In particular, $p_s < p_e$. See also the discussion of externalities in Section 2.1.

the other customers, he does not normally take these extra costs into account when deciding to join. Society, on the other hand, does. Hence, a customer who selfishly decides to join might be refused entry by a social planner. This leads to the inequality, $p_s < p_e$.

Left to decide for themselves, customers behave in a socially optimal way if they do not generate negative externalities. Since this is usually not the case in a queueing environment, something needs to be done to induce customers to behave in this way. For example, customers can be made to internalize their externalities (i.e., bear them by themselves). This is exactly what is achieved by the classic regulation scheme of Pigovian tax (Pigou 1920), a tax set equal to the negative externalities each customer imposes on others when the socially optimal strategy is used by all. This logic motivates the three regulation schemes that are discussed in Section 3.1 below. The implementation of the Pigovian tax has two major drawbacks. First, money transfers require collection systems, which usually are costly to maintain. Second, in order to impose the right fee, a central planner usually needs to know the utility structure of all customers as well as the structure of the market, and, in particular, the values of the parameters C and R, as well as the rates λ and[4] μ. These drawbacks motivate us to look for alternative regulation schemes.

3. Regulation Schemes

We refer to a regulation scheme as a set of rules, administrated by a central planner, under which the customers' equilibrium behavior is socially optimal. In other words, we deal with mechanism designs that elicit socially optimal behavior from selfish customers.

3.1. Existing Regulation Schemes

There are three mechanisms we are aware of from the literature that regulate the arrival rate to an unobservable M/M/1 queueing system, namely, schemes under which, when customers are left to decide for themselves, the resulting equilibria induce the socially optimal arrival rate. They are: (1) paying a flat entry fee, (2) drafting a contract under which customers who join pay in accordance with a function of a to-be-realized random variable,[5] and (3) asking customers to pay for a preemptive priority parameter of their choice if they decide to join. We next brief on each of these approaches.

1. **Flat entry fee.** An entry fee-based regulation method for the unobservable M/M/1 model was first introduced in Edelson and Hilderbrand (1975). It is clear (see (1)) that p_e is monotone

[4] One exception is Hassin's priority purchasing scheme described in Hassin (1995) and discussed further in Section 3.1 here.

[5] Technically, the former scheme is a special (in fact, trivial) case of the latter.

decreasing with R. Moreover, for R small enough (as C/μ), $p_e = 0$. Hence, there exists a service valuation, call it $R - T$, such that

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

In other words, the imposition of the appropriate entry toll, denoted here by T_s, results in an equilibrium joining rate that coincides with the socially optimal one. It is simple to check that $T_s = R - \sqrt{CR/\mu}$. Moreover,

$$T_s = \frac{Cp_s\rho}{\mu(1 - p_s\rho)^2}. \tag{8}$$

Comparing this with (7), we conclude that the regulating flat entry fee coincides with the externalities that one who joins inflicts on others when the arrival rate is the socially optimal one.

2. **Contracts.** The following scheme is suggested in Haviv (2014). Customers arriving at the queue are first given the option to balk. If they join they sign a contract that says, "I will pay $h(X)$," where X is a random variable whose realization is not known to the customer when deciding and $h(x)$ is some function. Clearly, if $\mathrm{E}(h(X)) = T_s$ (see (8)), then such a scheme leads to the socially optimal joining rate. If X is a random variable that has nothing to do with the queue (for example, a value drawn by a random number generator), this may look artificial: one may just as well impose the fixed entry fee of T_s. Such schemes might be made more appealing by relating X to the whereabouts of the customer who joins. For example, in Kelly (1991) it is suggested to charge $C\mu W^2/2 - CW$, where W is the (random) time spent in the system. Another approach suggested in Haviv (2014) is as follows. Let Y be the (random) externalities that the customer imposes on others. If it is possible to observe a random variable X, then defining $h(X)$ as the expected externalities given X (denoted by $\mathrm{E}(Y|X)$) will do since $\mathrm{E}(h(X)) = \mathrm{E}(\mathrm{E}(Y|X)) = \mathrm{E}(Y) = T_s$. In Haviv (2014), the conditional expected externalities are derived given four options for X: (1) time spent in the system, (2) queue length upon arrival, (3) queue length upon departure, and (4) customer's service time. The last item in this list is the most appealing as it is based on a characteristic of only the customer himself. For more on expected externalities conditioning on the length of the service requirement, see Haviv and Ritov (1998).

3. **Purchasing priority.** Hassin (1995) (see also Hassin and Haviv 2003, pp. 96–98) considers the following auctioning mechanism. Customers decide whether to join a queue or balk. A joining customer pays a nonnegative amount of his choice and enjoys preemptive priority over customers who pay less, regardless of the difference between their payments.[6] From the social point of view only the resulting equilibrium joining rate matters, and this, as it turns out, equals $p_s\lambda$.

[6] For a model where priority is proportional to the payment see Haviv and van der Wal (1997).

3.2. Desired Properties of Regulation Schemes

We define the following five desired properties that are used to classify the regulation schemes described in this paper, both the existing schemes and our new ones. In particular, we check which of these of properties are satisfied by each scheme.

Property 1 *Any customer is free to join or balk. If he joins, he will be served in a finite period of time.*

This property is a key property. Without it one can simply use the following obvious and usually undesired scheme: deny any arriving customer access to the queue with probability $1 - p_s$. In other words, a central planner decides who joins and who does not.[7]

Property 2 *The queueing regime is work-conserving.*

A queueing regime is work-conserving if the server is never idle when the queue is not empty. Put differently, the total amount of work left in the system, reflecting the sum of unfinished service among customers in the system, coincides with the similar value in the case of a FCFS regime. Regulation schemes that possess this property do not give up on efficiency in order to elicit socially better behavior. For example, schemes that are based on delay tactics do not possess this property.

Property 3 *The scheme does not involve money transfers.*

As already noted by Leeman (1964), implementation of regulation schemes that involve money transfers require the setup and maintenance of a collection system, which typically comes with costs. In addition to that, there are variable costs associated with each money transfer, such as transaction costs. Hence, schemes that do not involve transfers, if available, are preferable, especially if these constant and variable collection costs are relatively high compared to the social benefit of regulation.

Property 4 *The rules of the scheme are insensitive to the rate parameters μ and λ.*

Property 5 *The rules of the scheme are insensitive to the monetary parameters C and R.*

Obviously, in order to implement schemes that are sensitive to model parameters, one should know their values. Moreover, implementation of such schemes using wrong model parameter values may result in a lower social welfare than in an unregulated system. Hence, schemes that are insensitive to model parameters have an advantage over sensitive schemes. Furthermore, by using

[7] In addition to being undesired, this scheme does not possess properties 4 and 5 below.

insensitive schemes one can avoid estimation issues associated with the M/M/1 model (see, e.g., Schruben and Kulkarni 1982).

Consider the three schemes described in Section 3.1 above. The first two do not satisfy properties 3–5 while the third scheme satisfies all but property 3. Indeed, the flat entry fee scheme obviously does not satisfy property 3. As T_s is sensitive to all the model parameters, it also does not satisfy properties 4 and 5. For the contract-based schemes, property 3 is not satisfied as it involves money transfers. For properties 4 and 5, the description above does not exclude the possibility that there exist a random variable X and a function h such that $\mathrm{E}(h(X)) = T_s$, where h is insensitive to all the model parameters. Nevertheless, we are not aware of such a regulating contract in the existing literature and its existence is an open question. Finally, the purchasing priority scheme satisfies all the properties we are after except for property 3, as it is insensitive to all the model parameters but involves money transfers.

To the best of our knowledge, the existing literature contains no regulation scheme for this model that possesses all five above-mentioned properties. The main purpose of this paper is to suggest one. Furthermore, we suggest some variations of it, as well as additional schemes based on randomization, which do not possess one or more of properties 3–5, but nevertheless merit consideration.

3.3. The Preemptive Random Priority (PRP) Scheme

Suppose that customers (independently) draw a random preemptive priority parameter U uniformly distributed in the unit interval. The lower the value of U is, the higher the priority level is. In particular, a customer with parameter u preempts the service of a customer with parameter v, when $u < v$, if the former arrives and sees the latter being served. Moreover, the next customer to get service after service completion is the one with the lowest priority parameter in the queue.

Consider a customer who draws the parameter u. If all the customers whose priority is higher than his join, he becomes a standby customer in a similar system but with an arrival rate of $u\lambda$. From (5), his mean time in the system equals

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

while the unconditional mean waiting time of the customers with higher or equal priority remains the same as in the FCFS regime with joining probability u, i.e.,[8] $\frac{1}{\mu(1-u\rho)}$.

Customers, once informed of their (random) priority parameters, decide whether or not to join. It is assumed that it is common knowledge among customers that a preemptive random priority is the regime used and that all the customers ex-ante face the same situation. Under any strategy

[8] It is easy to verify that $\int_{x=0}^{u} \frac{1}{\mu(1-x\rho)^2}\frac{1}{u}\,dx = \frac{1}{\mu(1-u\rho)}$, as expected.

profile used by all the customers, the expected waiting time is monotone increasing in U and hence the individual best response is a threshold-based strategy: if U is less than or equal to some critical value, join. Otherwise, balk. Hence, a symmetric equilibrium is to join if and only if the priority parameter is less than or equal to u_e such that u_e is the unique solution of

$$R - \frac{C}{\mu(1-u\rho)^2} = 0 \tag{9}$$

for u. Recalling that U is uniformly distributed in the unit interval, u_e also equals the equilibrium joining probability. The following theorem claims that the scheme is a regulation scheme. Its proof follows immediately from comparing equations (3) and (9).

THEOREM 1. *The equilibrium joining probability under the Preemptive Random Priority (PRP) scheme and the socially optimal joining probability coincide. In the above notation,*

$$p_s = u_e.$$

Moreover, this scheme possesses all properties 1–5.

What the theorem says is that the preemptive random priority scheme is self-regulating in the sense that customers, when they behave in accordance with the resulting equilibrium, do in fact behave in a socially optimal way. An explanation of this phenomenon is as follows. The marginal customer, namely, the one to draw u_e, does not inflict any externalities (once all the other customers follow the equilibrium strategy) due to the nature of the queueing regime.[9] Hence, his (zero) utility due to joining coincides with the marginal social utility. In other words, since he is indifferent between joining and balking, the same is the case with society, and hence the probability of joining ought to be p_s.

REMARK 1. The result in Theorem 1 is invariant with respect to the distribution of U, the priority parameter, as long as it is continuous. Alternatively, it can be said that any continuous transformation of the original priority parameter U can be used as an alternative priority parameter. Moreover, instead of performing an actual lottery, one can use any (irrelevant) continuous heterogeneity of the customers (say, their biological age or their height) to assign priorities.

3.4. Variations of the PRP Scheme

The following two schemes are variations of the PRP scheme. Their advantage is that the decision making, from the customer's point of view, is more natural. This advantage comes at a price in the form of being sensitive to the model's parameters, i.e., not possessing properties 4 or 5.

[9] The preemption assumption does not change the overall mean waiting times. This fact does not hold under general service time distributions. See the discussion in Section 5.2 below.

1. **Random waiting time.** Instead of using the original priority parameter U, use its transformed value $\frac{1}{\mu(1-U\rho)^2}$ (see Remark 1). This function of U is the mean waiting time of a customer with priority parameter U in the original system, if all customers whose priority levels are higher join. Under this scheme (joining) customers are informed of their expected time in the system under this scenario, and hence the decision whether to join or balk is more natural. This scheme is sensitive to both rates μ and λ; i.e., it does not satisfy property 4.

2. **Binary random priority.** Customers draw their preemptive priority parameter I, where I is Bernoulli-distributed with parameter $1-p_s$. There is no need to specify what the rule of the scheme is within each class and it can be, for example, FCFS.[10] For those who get priority parameter 0 (a probability p_s event), joining is a dominant strategy since even if all customers join, their utility in case of joining is at least $R-\frac{C}{\mu(1-p_s\rho)}>0$. Now, for those who get priority parameter 1 (a probability $1-p_s$ event), their best response is not to join even if no one from their class joins (but those from the other class do join). This is due to the fact that in this (best-scenario) case, their utility equals $R-\frac{C}{\mu(1-p_s\rho)^2}=0$. The resulting joining probability is hence p_s, as required. Because p_s is sensitive to all four parameters, so is this scheme, and thus it does not satisfy properties 4 and 5.

3.5. The Random Entry Fee Scheme

Under this scheme, customers (independently) draw a random entry fee. Once informed of their random fee, customers decide whether to join (and pay the fee) or balk.

Theorem 2. *The random entry fee scheme is a regulation scheme if it is drawn from any distribution with CDF, denoted by $F(x)$, that satisfies*

$$F(T_s)=p_s, \tag{10}$$

where T_s is the optimal flat entry fee (see (8)) and p_s is the socially optimal joining probability (see (4)).

Proof. The equilibrium profile here is obviously a threshold strategy: join if and only if the randomly drawn entry fee is less than or equal to T_e, for some $T_e>0$. We next show that $T_e=T_s$ and hence, by the condition in (10), the equilibrium joining probability equals $F(T_s)=p_s$. In equilibrium, the net utility of a customer whose fee was drawn to equal the threshold fee T_e (and joins) should be equal to zero; i.e, T_e satisfies

$$R-T_e-\frac{C}{\mu(1-F(T_e)\rho)}=0. \tag{11}$$

[10] This choice leads to minimal variance in waiting times among those who join.

It is easy to verify that[11]

$$T_e = T_s,$$

with $F(T_e) = F(T_s) = p_s$, solves (11). □

This scheme does not satisfy properties 4–5 and also does not satisfy property 3 (lack of money transfers), except for special cases such as the following example of a binary lottery that satisfies a weak version of this property.

Remark 2. **Binary random entry fee.** Consider the following payment scheme: with probability p_s the fee is zero and with probability $1 - p_s$ it is any fee greater than T_s. Obviously, this distribution satisfies (10) and hence regulates the system. Interestingly, this version of the random entry fee scheme satisfies a weak version of property 3; i.e., the equilibrium behavior induced by the scheme does not come with money transfers. Yet, the option for money transfers is presented, but transfers are not part of the equilibrium path. Hence, variable costs due to money collection are avoided, although constant collection costs still exist as a collecting administration needs to be formed in order to make transfers possible and the scheme credible.

4. Summary

A succinct overview of all eight regulation schemes in this paper and their properties is given in Table 1. We add a column giving the resulting ratio between customer surplus and social welfare. Clearly this ratio is between zero and one, being equal to one when no money transfer takes place.

Scheme	Properties					$\frac{\text{Consumer surplus}}{\text{Social welfare}}$
	1	2	3	4	5	
Preemptive random priority	✓	✓	✓	✓	✓	1
Random waiting time	✓	✓	✓		✓	1
Binary random priority	✓	✓	✓			1
Random entry fee	✓	✓				$[0,1]$
Binary random entry fee[12]	✓	✓	✓			1
Flat entry fee	✓	✓				0
Contracts	✓	✓				0
Purchasing priority	✓	✓		✓	✓	0

Table 1 Summary of regulation schemes and their properties

[11] As expected, the threshold random fee coincides with the flat entry fee, which can be looked at as a degenerate example of a random fee.

[12] Possesses a weak version of Property 3.

Figure 1 summarizes the equilibria, social optimization, and some of the regulation methods mentioned in this paper. The curve labeled CW_{FCFS} represents the *mean* waiting cost under FCFS without any regulation as a function of the joining probability p. From the selfish point of view, the joining probability increases (respectively, decreases) as long as CW_{FCFS} is less than (respectively, greater than) R, and hence the intersection point of this function with the horizontal line R, where $p = p_e$, is the equilibrium point.

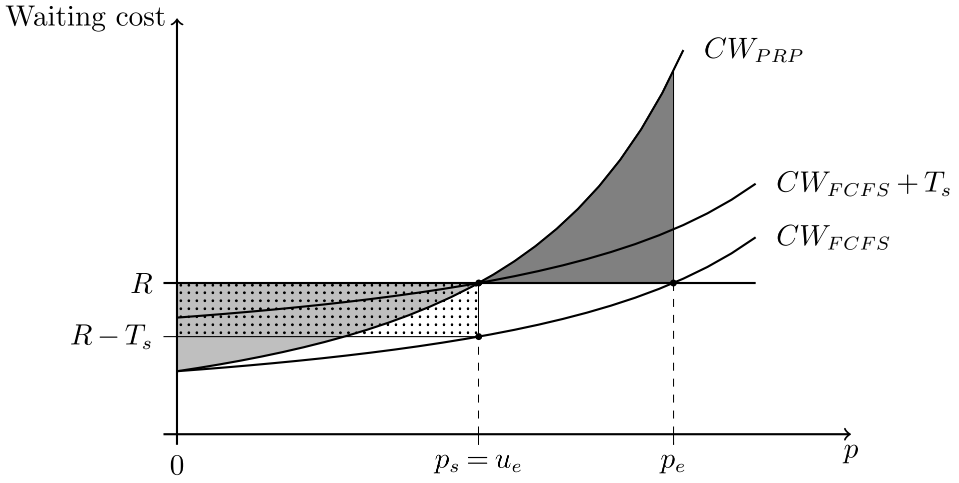

Figure 1 Waiting costs under FCFS and PRP and the corresponding equilibria

The curve labeled CW_{PRP} represents the waiting cost of a customer with priority parameter p, in the preemptive random priority system if all customers with higher priority levels join. Such a customer joins as long as $CW_{PRP} < R$, and hence the equilibrium point of this system is the intersection point of the CW_{PRP} curve with R, which takes place where $p = u_e$. This curve also represents the waiting cost of a customer in Hassin's model (Hassin 1995), whose payment is such that the proportion of customers who pay for a higher priority level is p. Therefore, the socially optimal probability is obtained at the intersection of this curve with the horizontal line R, namely, where $p = p_s$.

Consider a system with joining probability p. The difference between CW_{PRP} and CW_{FCFS} at p corresponds to the expected externalities that a customer who joins this system inflicts on the other customers. In particular, this difference at $p = p_s$ gives us T_s, the optimal flat entry fee (or the expected payment under any optimal contract). If this entry fee is charged, then the mean cost (waiting costs plus fee) is $CW_{FCFS} + T_s$ and the resulting equilibrium joining probability is p_s, as desired. As can be seen, the equilibrium joining rates of all regulation methods coincide with the socially optimal one.

Another observation that stems from this figure is as follows. The marginal utility due to an additional arrival to a system with a joining probability of p is the difference between R and the curve labeled CW_{PRP}. Therefore, the resulting mean utility of such a system is the area between these curves from zero to p. In the unregulated system, where p_e is the joining probability (and where the mean utility equals zero), the mean utility is also the light gray shaded area minus the dark gray shaded area, and hence the two areas are equal. In a system with the socially optimal joining probability, the mean utility is the light gray shaded area. Under all the regulation methods based on random priority that are stated in this paper, this social welfare also equals the consumer surplus as no transfer of money takes place. Conversely, in the case where a flat fee is introduced, subtracting the mean payment $T_s p_s$ (the dotted rectangle) from this mean utility yields the consumer surplus, which, as said before, equals zero. This means that the two areas, the light gray shaded one and the dotted rectangle, are equal. Note that under a strict random entry fee scheme, some of the customers pay less than T_s and hence the consumer surplus is strictly positive or even equals the social welfare, as in the binary lottery case.

In Hassin's model (Hassin 1995), the difference between R and the CW_{PRP} curve at p is the payment that the marginal customer at p pays. Therefore, the light gray shaded area is the mean payment in this case. As said before, this area is exactly the mean utility and the customers end up with a zero consumer surplus.

5. Discussion

5.1. Practical Issues

The analysis in this paper considers the mathematical and economical aspects of queue management. However, when dealing with humans in queueing contexts, some other aspects should be considered too as a waiting line can be seen as a "miniature social system" Mann (1969). In the normative contexts of queues, it seems that the most common, or even the natural, way is to grant service according to order of arrival. This idea is deeply ingrained in Western culture; see, e.g., Hall (1959) p. 201. For more on the game-analytic approach to norms in queues see Allon and Hanany (2012) and Yang et al. (2016). The PRP regime, among other socially optimal regimes suggested in the queueing literature, involves deviation from the FCFS norm. An additional feature of the PRP regime that might look revolutionary at first sight is that priority is granted by lottery. However, prioritizing by lottery is not a new idea. For example, since 1995, a limited number of U.S. permanent residence visas are granted under the Diversity Immigrant Visa lottery program each year "Green Card through the Diversity Immigrant Visa Program," (2016). Another recent example is the Buyer's Price housing assistance program in Israel, under which a limited number of families, selected by lottery, get a significant discount on buying their first house "Mechir Lamishtaken," (2016). Naturally, such lottery-based programs are often criticized as being unfair.

These two normative issues, like others that may arise, should be weighed when considering the implementation of the PRP scheme. In fact, the question that should be asked is, are we willing to adjust to new norms for the benefit of all? The answer of course depends on the time, the place, and other features of the specific system under consideration. Human history tells us that norms are flexible and it is certainly possible that the new regimes suggested in this paper, if implemented sagely, may even become the new norms in some situations.

There is one practical issue that was also addressed in Hassin (1985), which should be considered here as well. This is the issue of customers who may not be happy with their drawn random priority and hence may reappear shortly (at no cost). By doing so, they disguise themselves as new customers, hoping to draw a better priority level. Such behavior should be administratively prohibited or made unworthy. This can be easily done in service systems that require customers to be identified (e.g., in banks, hospitals, or call centers), as long as identification takes place upon arrival.

The implementation of the PRP scheme is possible only if preemptions are possible and come without any loss of work. Nevertheless, in the following variation of the M/M/1 model, a similar scheme is implementable without preemptions. Suppose that service is completed, and only then is it decided who among those in line shall be the one to receive it. This can be the case when service is standard and not customized. For example, one can think of a short-order cook who flips a burger and then decides whom to serve it to. The queue-length process and mean waiting time here coincide with those in the ordinary M/M/1 case, as do the optimal joining probabilities. The equivalent to the PRP regime in this scenario is granting completed service in accordance with randomly drawn priority parameters such that the customer who receives a completed service is the one with the lowest priority parameter present in the queue. This is probabilistically equivalent to the PRP regime in the ordinary M/M/1 model described in Section 3.3 and hence it regulates the system. Indeed, in cases of standard service the use of this combined service regime and regulation scheme is recommended as preemption is not needed.

5.2. General Service Times

The assumption that service times follow an exponential distribution plays an essential role in the analysis above. This is due, in particular, to the fact that the mean time in the system does not vary with the queue regime. However, this is not the case in general, which adds another aspect to the social optimization issue in the form of a choice of service regime. Specifically, denote by $\mathcal{R}$ the set of available service regimes. The optimal social welfare is defined by

$$\max_{r\in\mathcal{R},\ p\in[0,1]}\{p\lambda(R-CW_r(p))\},$$

where $W_r(p)$, $r \in \mathcal{R}$, $p \in [0,1]$ is the mean waiting time under the service regime r and the joining probability p. Note that by the above definition, if for two regimes, r_1 and r_2, $W_{r_1}(p) < W_{r_2}(p)$ for all $p \in [0,1]$, then social welfare is suboptimal when r_2 is practiced regardless of the joining rate.

Consider the FCFS and the Last-Come First-Served with Preemption-Resume[13] (LCFS-PR) regimes. It is shown in Haviv (2016) that for all $0 \le p \le 1$, $W_{FCFS}(p) < W_{PRP}(p) < W_{LCFS-PR}(p)$ if and only if the coefficient of variation of service time is greater than one (the reverse inequalities hold if the coefficient of variation is less than one). From that we learn that in every case, the PRP regime is dominated by one of the two above-referenced regimes; i.e., PRP cannot be a socially optimal regime once these two are available. However, using the PRP regime ensures a strictly positive social welfare,[14] which is better than the zero social welfare gained in an unregulated system.

As indicated by the comparisons made in this paper, there is no other known regulation method but the PRP that possesses all five properties and preserves the service regime. As a consequence, one should decide whether to give up on some properties, or to use the PRP regime at the price of some reduction in social welfare. In terms of the coefficient of variation of the service times distribution, the greater the deviation from the exponential assumption is, the greater the reduction. In one direction, when the coefficient of variation goes to infinity (while keeping the mean service time constant), the optimal welfare under PRP (as well as under FCFS) goes to zero while welfare under LCFS-PR is not affected.[15] In the opposite direction, when the coefficient of variation equals zero (deterministic service times), we are able to show that the welfare under PRP compared with the optimal one under FCFS is lower by at most one-third as the following theorem states.

Theorem 3. *Consider the M/G/1 system with arrival rate λ and a service times distribution with first and second moments of μ^{-1} and $\overline{x^2}$, respectively.*

Then

$$\inf_{R>C/\mu, \overline{x^2} \ge \mu^{-2}} \left\{ \frac{\max_{p\in[0,1]}\{p\lambda(R - CW_{PRP}(p))\}}{\max_{p\in[0,1]}\{p\lambda(R - CW_{FCFS}(p))\}} \right\} = \frac{2}{3}.$$

Moreover, this infimum is achieved under deterministic service times (when $\overline{x^2} = \mu^{-2}$).

A proof is given in the e-companion to this paper.

[13] The last to arrive have preemptive priority over those who have arrived earlier. Customers might be preempted while in service; and when they return to service, it is resumed from the point where it was interrupted last.

[14] In equilibrium, regardless of the service times distribution, those who hold the threshold priority level have zero utility while all the others with higher priority enjoy strictly positive utility.

[15] Mean waiting time under LCFS-PR is a function only of the first moment of the service distribution. See, e.g., Haviv (2013), p. 63.

5.3. Heterogeneous Customers

The idea of regulation via priorities can be extended to models with heterogeneous customers. Consider for example the model suggested in Littlechild (1974), which extends the unobservable cost-reward M/M/1 model by assuming that customers are heterogeneous with respect to their service valuation. This is modeled by assuming that each customer's service valuation R is a continuous nonnegative random variable with tail probabilities $\bar{F}(r) := \mathrm{P}(R \geq r)$. Waiting costs are linear with (homogeneous) cost per unit time C and arrival and service rates are denoted by λ and μ, respectively. Assuming that realizations of service valuations are observable by the central planner,[16] social optimization is achieved by admitting customers with $R \geq r^*$ where

$$r^* := \arg\max_{r \geq 0} \left\{ \bar{F}(r) \left(\mathbb{E}(R | R \geq r) - \frac{C}{\mu - \bar{F}(r)\lambda} \right) \right\}. \tag{12}$$

Assume now that preemptive priority is given based on customers' service valuation (the higher the valuation, the higher the priority). Similarly to the analysis in Section 3.3, it can be shown that the equilibrium condition coincides with the first-order condition of (12); i.e., this scheme regulates the system and achieves the socially optimal joining strategy.

5.4. Observable Queue

Consider the observable version of this model; that is, customers observe the queue length prior to deciding whether to join or balk. Assume service is granted on a FCFS basis. As shown in Naor (1969), both self- and social-optimization strategies are threshold strategies; i.e., join if and only if the queue length is below some integer value. Denote these thresholds by n_s for the social one and by n_e for the equilibrium. Not surprisingly, $n_e \geq n_s$, and hence regulation schemes should be considered. One such scheme was introduced in Hassin (1985). Under this scheme customers are served immediately upon arrival but are pushed back by later arrivals (with preemption, if necessary). The customer's decision here is not whether to join or balk, but when to renege when too many customers are ahead of him in line. It turns out that under this scheme, the resulting equilibrium threshold equals the socially optimal threshold n_s. The rationale behind that is that those at the rear of the queue, i.e., those who consider reneging, inflict no externalities on others and hence their decision whether or not to renege coincides with that of society. Note that this regulation scheme possesses all properties 1–5. In fact, our PRP scheme can be looked at as the counterpart of Hassin's scheme, this time for the unobservable version.

[16] This might be the case when different valuations require different service types, e.g., selling goods of varying values.

5.5. Conclusions

Regulation schemes for the unobservable M/M/1 queueing model in the existing literature are based on imposing on customers an additional cost that equals the externalities they inflict on others, i.e., makes them internalize these externalities. This internalization requires money transfers, which might be too costly or hard to implement. Moreover, these methods are usually sensitive to the model parameters. Under the new schemes described in this paper, customers do not necessarily internalize the exact externalities they impose. Some of them internalize more and some of them less, in a way that motivates each one of them to follow the socially optimal behavior. More precisely, the amount of additional costs imposed on each customer are chosen randomly, in a way that a proportion of p_s of the customers pay less than their externalities, and all the other customers are asked to pay more (and hence balk). In this way, the joining probability is p_s, as required. Moreover, the novel PRP scheme suggested in this paper is the only known regulation scheme for this model with the advantages of being insensitive to all model parameters and requiring no money transfers.

Acknowledgments

We thank the department editor, the associate editor, and the three anonymous referees for their valuable comments that greatly improved this paper. We thank Rafael Hassin for his careful reading and insightful comments throughout the process leading to this paper. This research was partly supported by Israel Science Foundation grants no. 1319/11 and 511/15.

References

Allon, G. and E. Hanany (2012), “Cutting in line: Social norms in queues,” *Management Science*, **58**, 493–506.

Alperstein, H. (1988), “Optimal pricing for service facility offering a set of priority prices,” *Management Science*, **34**, 666–671.

Chen, H. and M. Frank (2001), “State dependent pricing with a queue,” *IIE Transactions*, **33**, 847–860.

Edelson, N. M. and D. K. Hilderbrand (1975), “Congestion tolls for Poisson queueing processes,” *Econometrica*, **43**, 81–92.

Hall, E. T. (1959), *The Silent Language*, Doubleday.

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

Hassin, R. (1995), “Decentralized regulation of a queue,” *Management Science,* **41**, 163–173.

Hassin, R. and M. Haviv (2003), *To Queue or not to Queue: Equilibrium Behaviour in Queueing Systems,* Kluwer.

Haviv, M. (2013), *Queues: A Course in Queueing Theorey*, Springer.

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

Haviv, M. (2016), "The performance of a single server queue with preemptive random priorities," *Performance Evaluation* (accepted manuscript).

Haviv, M. and B. Oz (2016a), "On externalities in M/G/1 queue and standby customers" (in preparation).

Haviv, M. and B. Oz (2016b), "Regulating an observable M/M/1 queue," *Operations Research Letters*, **44**, 196–198.

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

Haviv, M. and J. van der Wal (1997), "Equilibrium strategies for processor sharing and random queues with relative priorities," *Probability in the Engineering and Informational Sciences,* **11**, 403–412.

Kelly, F. P. (1991), "Network routing," *Philosophy Transactions of the Royal Society,* **A337**, 343–367.

Leeman, W. A. (1964), "The reduction of queues through the use of price," *Operations Research* **12**, 783-785.

Littlechild, S. C. (1974), "Optimal arrival rate in a simple queueing system," *International Journal of Production Research,* **12**, 391–397.

Mann, L. (1969), "Queue culture: The waiting line as a social system," *American Journal of Sociology,* **75**, 340–54.

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

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

Schruben, L and R. Kulkarni (1982), "Some consequences of estimating parameters for the M/M/1 queue," *Operations Research Letters*, **1**, 75–78.

Yang, L., Debo, L., and V. Gupta (2016), "Trading time in a congested environment," *Management Science*, Advance online publication. doi: 10.1287/mnsc.2016.2436

"Green Card Through the Diversity Immigrant Visa Program," U.S. Citizenship and Immigration Services. Accessed August 04, 2016. https://www.uscis.gov/green-card/other-ways-get-green-card/green-card-through-diversity-immigration-visa-program/green-card-through-diversity-immigrant-visa-program.

"Mechir Lamishtaken," Ministry of Construction and Housing. Accessed August 04, 2016. http://www.moch.gov.il/English/housing_assistance/mechir_lamishtaken/Pages
/mechir_lamishtaken.aspx#GovXParagraphTitle2.

Proof of Theorem 3

Theorem 3. *Consider the M/G/1 system with arrival rate λ and a service times distribution with first and second moments of μ^{-1} and $\overline{x^2}$, respectively. Then*

$$\inf_{R>C/\mu,\overline{x^2}\geq\mu^{-2}}\left\{\frac{\max_{p\in[0,1]}\{p\lambda(R-CW_{PRP}(p))\}}{\max_{p\in[0,1]}\{p\lambda(R-CW_{FCFS}(p))\}}\right\}=\frac{2}{3}.$$

Moreover, this infimum is achieved under deterministic service times (when $\overline{x^2}=\mu^{-2}$).

Proof. Assume without loss of generality that $C=1$ and $\mu=1$. Then

$$W_{PRP}(p)=\frac{1}{1-p\lambda}\frac{\overline{x^2}}{2}-\frac{\ln(1-p\lambda)}{p\lambda}(1-\frac{\overline{x^2}}{2})$$

and

$$W_{FCFS}(p)=\frac{\lambda\overline{x^2}}{2(1-p\lambda)}+1,$$

where the former is given in Haviv (2014) and the latter is the well-known Khinchine–Pollaczek formula. The first-order conditions of p for the social optimization problem can be simplified to

$$R-\frac{1}{(1-p\lambda)^2}+\frac{p\lambda(1-\overline{x^2}/2)}{(1-p\lambda)^2}=0 \tag{EC.1}$$

under the PRP regime, and to

$$R-\frac{1}{(1-p\lambda)^2}+\frac{p\lambda(1-\overline{x^2}/2)(2-p\lambda)}{(1-p\lambda)^2}=0 \tag{EC.2}$$

under the FCFS regime. Note that in the case where $\overline{x^2}=2\bar{x}=2$, as in the exponential case, both conditions coincide with (3).

Solving (EC.2) and (EC.1) for p yields

$$\max_{p\in[0,1]}\{p\lambda(R-W_{FCFS}(p))\}=(\sqrt{R}-1)^2+2\sqrt{R}-\sqrt{4R-(\overline{x^2}-2)(2R+\overline{x^2})}+\overline{x^2}-2$$

and

$$\begin{aligned}\max\nolimits_{p\in[0,1]}\{p\lambda(R-W_{PRP}(p))\}=\ &(\sqrt{R}-1)^2+2\sqrt{R}-\sqrt{4R+(\overline{x^2}/2-1)(4R-1+\overline{x^2}/2)}\\&-(\overline{x^2}/2-1)\left[\ln\left(\frac{\sqrt{4R+(\overline{x^2}/2-1)(4R-1+\overline{x^2}/2)}-(\overline{x^2}/2-1)}{2R}\right)-1\right].\end{aligned}$$

By using the above and some cumbersome algebra (which we omit) it can be shown that

$$\frac{\max_{p\in[0,1]}\{p\lambda(R-W_{PRP}(p))\}}{\max_{p\in[0,1]}\{p\lambda(R-W_{FCFS}(p))\}}$$

is monotone increasing with $\overline{x^2}$ for any $R>1$. Hence, the infimum is obtained when $\overline{x^2}=1$, i.e., deterministic service times. In that case, the ratio is simplified to

$$\frac{2R-\sqrt{8R+1}+\log\left(\frac{\sqrt{8R+1}+1}{4R}\right)+1}{2R-2\sqrt{2R-1}},$$

which is monotone increasing in R, and hence

$$\inf_{R>1,\overline{x^2}\geq 1}\left\{\frac{\max_{p\in[0,1]}\{p\lambda(R-W_{PRP}(p))\}}{\max_{p\in[0,1]}\{p\lambda(R-W_{FCFS}(p))\}}\right\}=\lim_{R\to 1}\frac{2R-\sqrt{8R+1}+\log\left(\frac{\sqrt{8R+1}+1}{4R}\right)+1}{2R-2\sqrt{2R-1}}=\frac{2}{3}.$$

□