האוניברסיטה העברית בירושלים

THE HEBREW UNIVERSITY OF JERUSALEM

SELF-REGULATION OF A QUEUE VIA RANDOM PRIORITIES

By

MOSHE HAVIV and BINYAMIN OZ

Discussion Paper # 674 (December 2014)

מרכז פדרמן לחקר הרציונליות

THE FEDERMANN CENTER FOR THE STUDY OF RATIONALITY

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

Self-regulation of a queue via random priorities

Moshe Haviv and Binyamin Oz
Department of Statistics
and Federmann Center for the Study of Rationality
The Hebrew University of Jerusalem
91905 Jerusalem
Israel

December 20, 2014

Abstract

We consider a memoryless unobservable single-server queue where customers are homogeneous with respect to their reward (due to service completion) and with respect to their cost per unit of time of waiting. Left to themselves, it is well known that in equilibrium they will join the queue at a rate that is higher than it is socially optimal. We show that if customers draw a random preemptive priority parameter prior to deciding whether or not to join, the resulting equilibrium joining rate coincides with the socially optimal one. We also introduce some variations of this regulation scheme and review a few existing schemes from the literature. We suggest a classification of all these schemes, based on a few key properties, and use it to compare our new schemes with the existing ones.

1 Introduction

We deal here with the following decision model, which first appeared in [1]. Customers seek service from a single server in front of which a waiting line can be formed. Service times are exponentially distributed with a mean value of μ^{-1}. The potential arrival rate is denoted by λ customers per unit of time.

Assume that $\lambda < \mu$.[1] This arrival process is assumed to be Poisson. In summary, we have an M/M/1 queue. Each customer values service by R and it costs him C per unit of time in the system (service included). Customers are assumed to be risk neutral and hence are interested in maximizing their mean monetary utility. Recall that $1/(\mu - \lambda)$ is the mean time in the system under any work-conserving and non anticipating [2] queue regime such as First-Come First-Served (FCFS), assuming all customers join.

Assume without loss of generality that not joining the queue comes with a zero reward. In order to avoid trivialities, assume that $R > C/\mu$ 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, not joining).

Suppose now that each customer can decide whether or not to join the queue (without any further information such as the queue length upon arrival or her service requirement). The Nash equilibrium solution concept says that all customers should join with a probability p_e, where p_e solves the equation

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

for p. In equilibrium, those who join, as well as those who do not join, end up (on average) with nothing, making zero the consumer surplus under the equilibrium joining rate. It is easy see that

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

Note that $\lambda p_e < \mu$, whether $\lambda < \mu$ or not. In fact, λp_e, the effective arrival rate in equilibrium, is not a function of λ.

Society, as a single entity, would be better off if it (or a central planner on its behalf) controlled the joining probability. It would then be p_s, where p_s is defined as

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

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

[2] By "non anticipating" we mean that the customer 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.

In words, p_s is the joining probability that maximizes the mean net social utility gained per unit of time. Of course, if for some reason the joining probability is larger (respectively, smaller) than p_s, the central planner would like less (respectively, more) customers to join. It is only under p_s that he is indifferent whether an individual customer joins or not. It is easy to see that the first-order condition of the central planner's problem is

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

and the solution is

$$p_s = \frac{\mu - \sqrt{\frac{C\mu}{R}}}{\lambda}. \tag{3}$$

Clearly, $p_s < p_e$.[3] Again, λp_s, the socially optimal effective arrival rate, is not a function of λ.

One may look for a way to regulate the system so that selfish customers join with a probability of p_s, rather than p_e, when they solve for a Nash equilibrium of modified (or mechanically designed) decision making. The purpose of this paper is to introduce some novel regulation schemes, review a few such existing schemes from the literature, and suggest a classification of all these schemes based on five properties to be defined in Section 2. We present our original mechanisms in Section 3, and we review the existing ones in Section 4.

2 Regulation schemes: Definition and classification

We refer to a regulation scheme as a set of rules, administrated by a central planner, under which the equilibrium behavior coincides with the socially optimal behavior. In other words, we deal with mechanism designs that elicit socially optimal behavior from customers who, from their own point of view, behave selfishly. Examples of such rules are determining a service regime and setting an entry fee.

[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 utility and hence 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 on externalities in Section 4.

We use the following five properties 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.

1. Any customer can choose to join or not to join, and in the former case, he will be served in a finite period time.
2. The queueing regime is work-conserving.
3. The equilibrium behavior does not involve money transfers.
4. The rules of the scheme are insensitive to the two rate parameters, μ and λ.
5. The rules of the scheme are insensitive to the two monetary parameters, C and R.

Remark 2.1 The first 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.[4]

To the best of our knowledge, there is no regulation scheme for this model that possesses all five of the above-mentioned properties. The main purpose of this paper is to suggest one. Furthermore, we suggest some variations of this scheme which do not possess one or more of properties 3–5, which also merit consideration.

3 Regulation schemes

Before we introduce our four novel regulation schemes we discuss some useful notations. We denote by *stand-by* a customer who receives service only when the server would otherwise be idle. Note that a stand-by customer is preempted by those who arrive while he is receiving service. When there are no other customers in the system, the stad-by customer's service is resumed from the point where it was interrupted. It is well known that the mean time in the system for a stand-by customer equals

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

$$\frac{1}{\mu(1-\rho)^2}, \tag{4}$$

where $\rho = \lambda/\mu$ is the traffic intensity or the server utilization level (see, e.g. [6], p.64).

Remark 3.1 Suppose that all customers believe that they are stand-by customers. In that case, the equilibrium joining probability would be equal to p_s (see equation (2)). Nevertheless, this scenario cannot be a result of a regulation method since all being stand-by customers cannot be common knowledge.

Remark 3.2 Consider the following variation of the M/M/1 model. Suppose service is completed, and only then is it decided who will be the next to receive service among those in line. One can think of a short-order cook who cooks a hamburger and then decides who to serve it to.[5] In this scenario, a stand-by customer receives the order if and only if he is the only one in the system when the burger is ready to serve. The queue length process here coincides with that in the FCFS case. Likewise, the waiting times of the stand-by customers in both systems coincide.

3.1 The preemptive random priority scheme

Suppose that customers (independently) draw a random preemptive priority parameter U, where U is uniformly distributed in the unit interval. Assume that the lower is the value of U, the higher is the priority level. In particular, a customer with parameter u preempts 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.

Remark 3.3 The above-described scenario is probabilistically equivalent to the scenario described in Remark 3.2 (the short-order cook) where the customer who receives the completed service is the one with the lowest priority parameter present in the queue.

[5] Due to the memoryless service distribution it does not matter if the cook works when the system is empty, as long as he disposes of the burger if upon completion of its cooking there is no one there to receive it.

Consider a customer who draws the parameter u. It is clear that if all the customers whose priority is higher than his join, he becomes a stand-by customer in a similar system but with an arrival rate of λu. From (4), we learn that his mean time in the system equals

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

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.[6],

$$\frac{1}{\mu(1-\rho u)}.$$

Also, assume that customers, once informed of their (random) priority parameters, need to decide whether or not to join (when it is common knowledge that a preemptive random priority is the regime used and that all face the same ex-ante situation). It is clear that under any strategy profile used by all customers, the individual best response is a threshold-based strategy: if U is less than or equal to some critical value, join. Otherwise, do not join. Hence, a symmetric Nash equilibrium will be to join if and only if the priority parameter is less than or equal to u_e such that u_e is the unique value that solves the equation

$$R - \frac{C}{\mu(1-\rho u)^2} = 0 \tag{5}$$

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 (2) and (5).

Theorem 3.1 *The equilibrium joining probability under the preemptive random priority (PRP) scheme and the socially optimal joining probability, coincide. In the above notation,*

$$p_s = u_e.$$

This scheme possesses all properties 1–5.

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

What the theorem says is that the preemptive random parameter is self-regulating in the sense that customers, when they behave in accordance with the resulting Nash 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 use the Nash equilibrium strategy) due to the nature of the queueing regime.[7] Hence, his (zero) utility due to joining coincides with that of society (which does not mind who joins and who does not, as long as the probability of joining across all arrivals is p_s). Since he is indifferent between joining and not joining, the same is the case with society, and hence the probability of joining ought to be p_s.

Remark 3.4 The result in Theorem 3.1 is invariant with respect to the distribution of U, the priority parameter, as long as the distribution 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.

Remark 3.5 Instead of performing an actual lottery here, one can use any (irrelevant) continuous heterogeneity of the customers (say, their biological age or their height) to assign priorities.

Remark 3.6 Note that the mean time in the system for a stand-by customer under the socially optimal joining rate equals

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

Remark 3.7 Consider the observable version of this model; that is, customers observe the queue length prior to deciding whether or not to join. As shown in [13], both self- and social-optimization strategies are threshold strategies; i.e., join if and only if the queue length is below some threshold. Denote these thresholds by n_s for the social one and by n_e for the selfish one. Not surprisingly, $n_e \geq n_s$, and hence regulation schemes should be considered. One such scheme was introduced in [3]. Under this scheme customers are served immediately upon arrival but are pushed back by later arrivals

[7] The preemption phenomenon does not change the overall mean waiting times. This fact does not hold under general service time distributions.

(with preemption, if necessary). The customers' decision here is not whether or not to join, but when to renege when too many customers are ahead of them in line. It turns out that under this scheme, the resulting threshold is equal to the socially optimal threshold n_s; i.e., this is a regulation scheme that possesses all properties 1–5.

3.2 The random waiting time schemes

The next two schemes are in fact variations of the PRP scheme. The advantage of these schemes is that decision making, from the customers' point of view, is more natural. This advantage comes with a price in the form of being sensitive to the model's parameters, i.e., not possessing properties 4 and/or 5.

The first scheme is as follows. Instead of using the original priority parameter U, use its transformation $\frac{1}{\mu(1-\rho U)^2}$ (see Remark 3.4). This function of U is the mean waiting time of a customer with priority parameter U in the original system, if all the customers whose priority is higher than his join. Under this scheme the customers are informed of their expected time in the system prior to deciding whether or not to join. Note that this scheme is sensitive to both rates μ and λ; i.e., it does not satisfy property 4.

The second scheme is as follows. The customers are informed of the value of $1/(\mu(1-\rho U)^2)$ only if $U \leq p_s$. Customers with a higher value of U are now informed of (a bit more than) R/C (see (6)). In this way all (joining and not joining) customers are informed of their expected waiting time if everyone behaves in accordance with the prescribed Nash equilibrium behavior. This scheme is sensitive to all four parameters, μ, λ, C, and R; i.e., it does not satisfy properties 4 and 5.

3.3 The binary random priority scheme

Consider one more variation of the above schemes, in which properties 4 and 5 are not satisfied. Customers draw their preemptive priority parameter I, but now I is Bernoulli-distributed with probability $(1 - p_s)$. There is no need to specify what the rule of the scheme is, e.g., FCFS, among those who belong to the same priority class.[8] For those who get priority parameter 0 (with

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

probability p_s), joining is a dominant strategy since even if all customers join, their utility in case of joining is at least $R - \frac{C}{\mu(1-\rho p_s)} > 0$. 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 $R - \frac{C}{\mu(1-\rho p_s)^2} = 0$. The resulting joining probability is hence p_s, as required.

3.4 The random entry fee scheme

Under this scheme, customers (independently) draw a random entry fee. Once informed of their random fee, customers then decide whether or not to join (and pay the fee).

Theorem 3.2 *The random entry fee scheme is a regulation scheme if it is drawn from any distribution with CDF, denoted by $F(x)$, that satisfies*

$$F\left(\frac{Cp_s\rho}{\mu(1-p_s\rho)^2}\right) = p_s.$$

Proof: The (symmetric) equilibrium profile here is obviously a threshold strategy: join if and only if the (random) fee is less than or equal to $T > 0$. The equilibrium joining probability is therefore $F(T)$. Under such a lottery, the equilibrium strategy is $T^* := \frac{Cp_s\rho}{\mu(1-p_s\rho)^2}$. To see this, observe that if all customers follow this threshold strategy, one best response is to join if and only if one's fee is less than or equal to T^*. Indeed, the net utility of a customer with a (random) fee x (assuming all customers adopt this strategy) is

$$R - x - \frac{C}{\mu(1-p_s\rho)}.$$

Some algebra shows that this net utility is greater than or equal to zero if and only if $x \leq T^*$, as required. ∎

This scheme does not satisfy properties 4–5. This holds true for property 3 (lack of money transfers), except for the following example of a binary lottery: with probability p_s the fee is zero and in the complementary case it is any fee that is greater than T^*. Interestingly, this scheme does satisfy property 3; i.e.,

it does not involve a money transfer under equilibrium behavior (although the possibility of a money transfer exists). Another special case of this scheme is simply a flat entry fee of T^*. The flat entry fee is discussed further in the next section.

Remark 3.8 Another point we would like to make here takes us outside the model. In reality, customers who are unhappy with their draw of the lottery have an incentive to balk and reappear shortly afterwards disguised as new customers. This should be disallowed and, in fact, with modern technology our suggestions can be enforced.

Remark 3.9 The assumption that service times follow an exponential distribution is essential here. In general, the mean time in the system is not invariant with the queue regime. In particular, when comparing this mean between the FCFS and the PRP cases, we get that the former is higher if and only if the coefficient of variation of service time is greater than or equal to 1. For this result and more on the M/G/1 PRP model, see [8].

4 Discussion

In this section we review first the concept of externalities, especially in the context of queues. Then we deal with three existing regulation methods. The first is based on a flat entry fee. The second is based on contracts that customers need to sign prior to entering the system, whereby they commit to pay in accordance with a to-be-realized random variable. The third is based on allowing customers to pay for priority. We conclude with a summary on the various mechanisms presented here.

4.1 Externalities

In the context of queues, the concept of stand-by customers is closely related to that of *externalities*. For a comprehensive discussion on this relation see [12]. The externalities that a tagged customer imposes on the other customers are the total added time that the other customers wait compared to how long they would have waited if that tagged customer had not arrived. In order to estimate this added time, we can make this tagged customer a stand-by customer. By doing so we do not change the aggregate waiting time of all customers (including the tagged customer himself) 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 borne himself anyhow. Hence, the increase of his waiting time,

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

equals the externalities that the tagged customer imposes on the other customers.

Remark 4.1 There is an alternative way to look at externalities. Tag a customer who joins an M/M/1 queue. Next, Compare the process of the number of customers in this queue with that in a similar simulated queue without the tagged customer. By definition, the simulated queue will have exactly the same arrival and service processes as the original one. The two queue couple for the first time when the simulated queue becomes empty for the first time. Moreover, throughout this period the original queue has exactly one more customer than the simulated one. Of course, it is not necessarily the same customer who is the additional one but it will always be the tagged customer if we declare him a stand-by customer. Yet, from the social point of view it does not matter who the extra customers are in this period. Hence, the length of this period can be seen as the marginal total waiting time added to society due to the joining of the tagged customer (which is invariant with the queue regime due to the memoryless property of the service distribution). By subtracting from this added time the regular customer's (mean) time in the system and we get the externalities that a tagged customer imposes on the other customers. Note that no externalities are imposed by a stand-by customer, while the externalities imposed by an arrival in the case of a non preemptive queue regime such as FCFS equal

$$\frac{1}{\mu(1-\rho)^2} - \frac{1}{\mu(1-\rho)} = \frac{\rho}{\mu(1-\rho)^2}.$$

The presence of externalities is the root cause of the difference between the selfish behavior of customers (when they are left to decide for themselves) and socially optimal behavior. Consider for example the "to queue or not to queue" type of decision making we deal with here. Although a customer who joins inflicts (negative) externalities on the other customers, he does not normally take these extra costs into account when deciding whether or not to join. Society, on the other hand, does. Hence, a customer who selfishly

decides to join might be refused entry by a social planner who takes into account the externalities too. Hence, $p_s < p_e$. Left to decide for themselves, customers will behave in a socially optimal way only if they do not generate any negative externalities (and their behavior does not affect the overall performance as is the case in an M/M/1 model). Since this is seldom the case, something needs to be done to make customers behave in a socially optimal way. In other words, customers must be made to internalize their externalities (i.e., bear them by themselves) in order to achieve their socially optimal behavior. This is exactly what is achieved by the preemptive random priority (PRP) scheme. Specifically, assume a threshold-based behavior; i.e., up to some parameter value u, customers join, and for larger values, they do not. The threshold customer, namely, the one who holds the priority level of u_e (and hence joins), inflicts no externalities: given this behavior from everyone, the customer's and society's utilities coincide. Hence, if a customer is indifferent between joining and not joining, so is society. Indeed, customers with a priority parameter lower than u_e are welcome to join, whereas customers with a higher parameter are rejected by society.

There are three mechanisms we know from the literature that lead to the regulation of the arrival rate, namely, schemes under which, when customers are left to decide for themselves, the resulting Nash equilibria lead to 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,[9] and (3) asking customers to pay for a preemptive priority parameter if they decide to join. We next elaborate on each of these mechanisms.

4.2 Flat entry fee

It is clear (see (1)) that p_e is monotone decreasing with R. Moreover, for R small enough (actually, $R \leq C/\mu$), $p_e = 0$. Hence, there exists a reward, call it $R - T$, such that

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

In other words, the imposition of the right entry toll, denoted here by T, results in an equilibrium joining rate that coincides with the socially optimal

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

one. This will regulate the system. It is a simple exercise to check that $T = R - \sqrt{CR/\mu}$. More importantly, as it turns out,

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

Comparing this with (7), we conclude that the regulating flat entry fee coincides with the externalities that one who joins this FCFS M/M/1 queue inflicts on others when the arrival rate is the socially optimal one.

4.3 Contracts

The following scheme is suggested in [7]. Customers who arrive at the queue are first given the option to opt out. If they join, they sign a contract that says: "I will pay $f(X)$," where X is a random number (whose realization is not known to the customer when deciding whether or not to join) and where $f(x)$ is some function. Clearly, if $\mathrm{E}(f(X)) = T$, 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. Such schemes might be more appealing in the case where X is related to the whereabout of the customer who joins. In particular, let Y be the externalities that the customer imposes on others. If it possible to observe a random variable X, then defining $f(X)$ as the expected externalities given X (denoted by $\mathrm{E}(Y|X)$), will do since $\mathrm{E}(f(X)) = \mathrm{E}(\mathrm{E}(Y|X)) = \mathrm{E}(Y) = T$. In [7], four options for X are suggested and the expected externalities given X are derived. They are:

1. **The time in the system.** Denote it by W. Then $\mathrm{E}(Y|W) = (\sqrt{CR\mu} - C)W$. In particular, it is a linear function in W.
2. **The queue length upon arrival.** Denote it by L_a. Then $\mathrm{E}(Y|L_a) = (\sqrt{CR/\mu} - C/\mu)L_a$. In particular, it is a linear function in L_a.
3. **The queue length upon departure.** Denote it by L_d. Then $\mathrm{E}(Y|L_d) = \sqrt{CR/\mu}L_d$. In particular, it is a linear function in L_d.
4. **The customer's service time.** Denote it by S. Then, $\mathrm{E}(Y|S) = \frac{\mu C}{2}(\sqrt{R\mu/C} - 1)S^2 + C(\sqrt{R\mu/C} - 1)^2 S$. In particular, it is a quadratic function in[10] S.

[10] For more on $\mathrm{E}(Y|S)$ see [9].

Finally, schemes other than those based on expected conditional externalities are possible, as long as the requirement $\mathrm{E}(f(X)) = T$ is maintained. Kelly [11] (see also [7]) suggests two such schemes. They are $C\mu W^2/2 - CW$ and $C/(2\mu)(L_a^2 + L_a)$. Their advantage is that they are invariant with respect to R. In other words, they need not change when R chances; in fact, they can be imposed without the central planner even knowing the value of R. Note that the latter scheme is implementable even without knowing the values for C and μ: only their ratio is required. Note that all versions of this scheme do not satisfy properties 3–5.

4.4 Purchasing priority

Hassin [4] (see also [5], pp. 96–98) considers the following scenario. Customers need to decide whether they want to join a queue or not. In the former case, they need to pay a nonnegative amount of their choice. A customer who pays more enjoys preemptive priority over a customer who pays less, regardless of the difference between their payments.[11] Hassin indeed finds the proportion of those who join and the continuous distribution over the amount they pay to be in equilibrium. Yet, from the social point of view only the equilibrium joining rate matters. Hassin then shows that this joining rate equals λp_s. In other words, making all the options for paying for preemptive priority available regulates the system. Hassin also shows that in equilibrium each customer pay the externalities that he inflicts on the others, provided that all customers behave in accordance with the equilibrium profile.

Hassin's argument goes as follows. In the mixed strategy stating how much to pay for priority, there cannot be an atom: otherwise each customer has an incentive to pay a bit more than those at the atom. This infinitesimal extra payment will lead to a quantum of gain, thereby refuting the conjecture that the stated strategy is an equilibrium profile. Hence, the mixing result in continuous density. This density does not come with gaps and moreover it commences at zero payment. Indeed, otherwise the customer who should pay the lowest value would be better off shifting to zero and saving a quantum without losing anything in terms of priority. Finally, let us consider a customer who pays zero. He is a stand-by customer among all those who receive service. His utility, as everyone else's, is zero. Hence, the joining rate

[11] For a model where priority is proportional to the payment see [10].

is as prescribed by social optimization. Note that this scheme satisfies all the properties we are after except property 3.

4.5 Summary

A succinct overview of all eight regulation schemes in this paper, their properties, and the resulting ratio between customer surplus and social utility is given in Table 1.

Scheme	Properties					$\frac{\text{Consumer surplus}}{\text{Social utility}}$
	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	✓	✓	✓			1
Flat entry fee	✓	✓				0
Contracts	✓	✓				0
Purchasing priority	✓	✓		✓	✓	0

Table 1: Summery of regulation schemes and their properties

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 regulation as a function of the joining probability p. From the selfish point of view, the joining probability will increase (respectively, decrease) 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.

The curve labeled CW_{PRP} represents the waiting cost in a preemptive random priority system, of a customer with priority parameter p, if all customers with higher priority parameters join. Such a customer will join 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, where $p = u_e$. This curve also represents the waiting cost of a customer in Hassin's model in [4], 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

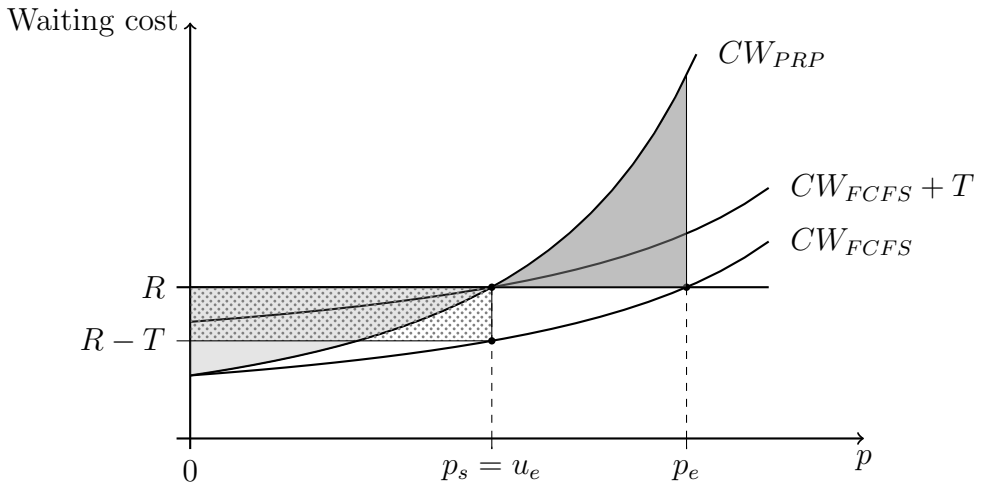

Figure 1: Waiting costs under FCFS and PRP and the corresponding equilibria

intersection of this curve with the horizontal line R, namely, where $p = p_s$.

Moreover, the difference between CW_{PRP} and CW_{FCFS} at some given point p is the mean externalities that a customer who decides to join a system with joining probability p inflicts on the other customers. In particular, this difference at $p = p_s$ gives us T, 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 $CW_{FCFS} + T$ and the resulting equilibrium joining probability is p_s, as desired. As can be seen, the equilibrium joining rates of all four 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 joining probability p is the difference between R and the curve labeled CW_{PRP}. Therefore, the mean utility of a system with a joining probability of p is the area between these curves from zero to p. In the unregulated FCFS 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. Note that under all the regulation methods based on random priority in this paper, this social utility also equals the consumer surplus. However, in the case where

the flat fee is introduced, subtracting the mean payment Tp_s (the dotted rectangle) from this mean utility gives 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 and hence the consumer surplus is strictly positive or even equals the social utility, as in the binary lottery case.

In Hassin's model, 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 system. As said before, this area is exactly the mean utility and the customers end up with zero consumer surplus.

Acknowledgement

This research was partly supported by Israel Science Foundation grant no. 1319/11.

References

[1] Edleson, N. M. and D. K. Hildebrand (1975), "Congestion tolls for Poisson queueing processes," *Econometrica*, **43**, 81–92.

[2] Glazer, A. and R. Hassin (1986), "Stable priority purchasing in queues," *Operations Reserch Letters,* **4**, 285–288.

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

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

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

[6] Haviv, M. (2013), *Queues - A Course in Queueing Theorey*, Springer.

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

[8] Haviv, M. (2014), "The M/G/1 queueing model with preemptive random priorities," *Proceedings of VALUETOOLS 2014*.

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

[10] 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.

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

[12] Haviv, M. and B. Oz (2014), "On externalities in M/G/1 queue and stand-by customers," (in preparation)

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