

Decentralized Regulation of a Queue

Refael Hassin

*Department of Statistics and Operations Research, School of Mathematical Sciences,
Tel Aviv University, Tel Aviv 69978, Israel*

A bidding mechanism for determining priorities in a service system is analyzed. It is shown that when all customers have the same exponential service demand, this mechanism induces both the socially optimal arrival process and the service order. It is also shown that the profit-maximizing service rate in this model is smaller than or equal to the socially optimal one.

(Priorities in Queues; Regulation of a Queue; Optimal Rate of Service)

1. Introduction

In some queues an arriving customer observes the queue length before deciding whether to join the queue or not. We refer to such queues as *queues with balking*. In other cases, the queue length cannot be observed, and customers base their decision to join the queue on statistical data possibly gathered from past experience. We refer to such queues as *queues without balking*. Naor (1969) observed that in queues with balking the individual's decision deviates from the socially preferred one. This gap is caused by the existence of external effects associated with the act of joining the queue, namely, future customers may have to wait longer. The individual's objective does not take these externalities into consideration, in contrast to the social objective. Edelson and Hildebrand (1975) showed that a similar phenomenon exists in queues without balking. There, if the demand for service is large, the system may become too congested if all this demand has to be served. In such a case, the arrival rate reaches an equilibrium size such that customers are indifferent between joining the queue or not. Because of the negative external effects mentioned above, this equilibrium rate is larger than the socially desired one. (See Mills 1981 for a more general discussion of this issue.)

As a consequence of the suboptimality of individual decisions, there is much interest in finding methods for regulating the arrival process. Naor's model and results have been extended in several papers. Comprehensive surveys are given by Stidham (1985) and Mendelson and Whang (1990). A recent paper is Haviv (1991).

Several methods have been suggested to regulate the arrival rate to a queueing system. The most straightforward approach is to administratively forbid entrance when this is socially desired. Alternatively, one may impose admission fees to deter customers from joining long queues. Naor showed, however, that in the balking model, the profit-maximizing toll is too high relative to the socially optimal one. Therefore, a correct toll must be computed and administratively imposed. Other mechanisms have been proposed to regulate the arrival rate, and most of them involve estimation by the system administrator of certain parameters necessary to compute fees imposed on customers. This is the case with Edelson and Hildebrand's admission toll and two-part tariff, and Mendelson and Whang's (1990) priority price. An exception is Hassin's (1985) first-come last-served discipline in a queue with balking. This discipline is *self-regulating* in the sense that the system administrator need not compute or install any fees. The system is regulated by the equilibrium behavior of customers who adjust their actions according to their experience. In some cases, the queue administrator can choose (by controlling information about the queue length) between the balk and no-balk regimes. The effect of such an option on profits and social welfare is analyzed by Hassin (1986).

When potential customers come from a heterogeneous population and the priority they obtain depends on their characteristics, there is a possibility that they will increase their welfare by giving incorrect information. The discipline suggested by Mendelson and

Whang is *decentralized* in the sense that each arrival is free to choose his own priority level. The specific prices they suggest are also *incentive-compatible* in the sense that in equilibrium customers will buy priorities in the socially desired way. Similar ideas were suggested by Marchand (1974) and Dolan (1978) in less general models where the arrival rate is independent of the prices.

There is another source for possible suboptimality in service systems, and this is the service rate, when it is a decision variable chosen by a profit maximizing server. Grassmann (1979) derived conditions for social optimality of the service rate in a general queueing model. However, Grassmann assumed that the service rate does not affect the arrival rate to the system. Edelson and Hildebrand proved that in their no-balking model the server will select both the socially optimal toll and service rate. However, they observed (see Footnote 1 in their paper) that in a model with balking the rate chosen by a profit maximizer will in general not be socially optimal. Dewan and Mendelson (1990) derive (social) optimality conditions when the service rate and admission fees are jointly determined.

The subject of this paper is the regulation of the arrival process in a queue with exponential service and without balking by a decentralized self-regulating mechanism: An incoming customer offers a payment for purchasing priority, and then customers with higher payments receive higher priority, with the possibility of service preemption. (Similar systems, with a different focus, were previously analyzed by Kleinrock (1967), Lui (1985), and Glazer and Hassin (1986). We first follow Edelson and Hildebrand and analyze a model with identical customers. We show that the socially optimal arrival rate is attained. We then follow Mendelson and Whang and analyze the model when customers differ by their service valuation and waiting costs. We show that, in this case, the (socially) optimal arrival rate with the optimal composition of customers is attained, and that customers will also be served in the optimal order.

Myrdal (1968), referring to queueing systems in which customers pay bribes to obtain priorities, suggested the possibility that the server will be motivated to slow down service in order to increase profits. Lui (1985) argued that Myrdal's hypothesis is not always true. For example, if increasing the rate of service is

costly to the server, then without bribe the server has no incentive for supplying fast service. Indeed, since the optimal rate is zero, bribes induce faster service.

In contrast to Lui's point of view, we compare the service rate chosen by a profit maximizer to the socially optimal one. We show that from this point of view, Myrdal's hypothesis is correct: The service rate chosen by a profit-maximizing server is either smaller than or equal to the socially optimal one.

Samuelson (1985) considered allocation of a unit of a good through auctioning in which applicants bear bid-preparation costs. It is assumed that the potential applicants' valuations of the good are random variables from a common distribution function, and each realization is known to the specific potential applicant only. When an individual decides to participate in the auction, he ignores the fact that by doing so he may prevent another customer from receiving the good. Because of this external effect his decision to apply may differ from the social optimal decision. Under the assumption that applicants value the good differently, there exists an equilibrium solution which is incentive-compatible in the sense that those who value the good highly will participate in the auction, and the higher is their valuation of the good the higher is their bid. Samuelson showed that the equilibrium participation in the auction is socially optimal. This result, that can be extended to the case where several identical items are distributed, adds to other results demonstrating that auctions can be used to discriminate optimally among individuals from a heterogeneous population.

Before presenting our main results, on the queueing model we treat in §2 a variation of Samuelson's model. We assume that several (possibly nonidentical) items are allocated to individuals from a *homogeneous population*. Since we assume that potential applicants make their decisions independently, there is no way to guarantee that the optimal participation will be achieved. This is in contrast to heterogeneous models like Samuelson's. However, in equilibrium the socially optimal *probability* of participation is achieved. Properties proved for this model are used in §3 to prove the optimality of the arrival rate in the queueing model with identical customers. Questions regarding the rate of service are analyzed in §4 and 5. The optimality of the proposed discipline in a queueing model with hetero-

geneous population is demonstrated in §6. The final section contains some concluding remarks.

2. Regulating the Number of Applications

Individuals from a homogeneous population independently decide whether to apply, wishing to obtain a single unit from a collection of (possibly nonidentical) items. A positive preparation cost is incurred on each applicant. To exclude trivial cases we assume that the number of items of maximum value is smaller than the number of potential applicants. We assume that social welfare is the total value of items that are allocated minus the sum of preparation costs of all individuals. A probability of applying is said to be *optimal* if it maximizes social welfare. Since applicants are identical, the way items are allocated among them is immaterial as long as we take care that highly valued items are distributed first. The only variable to be controlled is therefore the number of applications or, since individuals make their decisions independently, the probability that a randomly chosen individual will decide to apply.

We assume that each applicant must offer with his application a nonnegative payment ("bid," "bribe"), and items are distributed giving priority to higher offers: The most highly valued item is given to the applicant who offered the highest amount, the second item to the second highest offer, and so on. Ties are broken in some random way. The results below hold for two models:

Model I: Payments are made only by those applicants who obtain a unit of the good.

Model II: Payments are made by all applicants and are not returned, even to those who do not obtain any item.

In both cases the offer is irrevocable, and according to our definition of social welfare, the payments are considered as a transfer of income that does not affect social welfare.

We assume that each individual knows the total number of potential applicants. Moreover, no individual knows how many apply and how much they offer, but each knows the statistical distributions of these random variables. Below we characterize these distributions in *equilibrium*, that is, under the assumption that given that these distributions hold, every individual maximizes

his own welfare by sticking to his policy of whether to apply and how much to offer.

Since we assume that all customers are identical, in equilibrium the expected welfare associated with each of the possible payments must be identical. Otherwise, payments with low expected welfare do not optimize the individual's objective. Hence, each potential applicant must have identical expected welfare.

Let X be the random variable denoting the amount offered by a randomly chosen applicant. Let B denote an equilibrium cumulative probability distribution of X . Then:

$$B \text{ is continuous,} \quad (2.1)$$

B is strictly increasing on an interval $[0, a]$,

$$\text{where } a > 0 \text{ and } B(a) = 1, \quad (2.2)$$

and

the equilibrium probability of application

$$\text{is optimal.} \quad (2.3)$$

Equation (2.1) follows since a discontinuity in B at $X = x$ means that with a positive probability there is already an applicant who offered exactly x . Therefore an applicant's expected welfare will increase if he offers $x + dx$ rather than x (and increases his chances to obtain an item or a better item with positive probability). This contradicts the condition that in equilibrium all applicants have the same expected benefit.

Equation (2.2) follows since if $B(x)$ is constant for $b \leq x \leq d$ where $0 \leq b < d$, but increases for $x > d$, then an applicant who offers b instead of d reduces his expenses without increasing the risk of getting a less valuable item or not getting an item at all. This again contradicts the equilibrium assumption. Positivity of a follows since we excluded the possibility that the number of items of maximum value exceeds the number of potential applicants.

To prove (2.3), note again that in equilibrium potential applicants all have equal expected welfare. From (2.2), an individual applies if and only if he is ready to apply with $x = 0$. From (2.1) there is no probability mass at $x = 0$ (or at any other amount) so that in this case the applicant is certain to pay the lowest amount and to obtain the lowest priority while allocation takes

place. In that case he imposes no externalities on others. In other words, he will get an item if and only if all other applicants get better (or equal) items. Therefore, his decision to apply results from considerations identical to those that maximize social welfare. We conclude that an applicant applies if and only if it is socially desired that he do so.

A numerical example may clarify the above discussion. Consider Model II with two items of values \$10,000 and \$1000. Assume that the application preparation cost is \$300 and that customers are risk neutral. Suppose first that there are only two potential applicants. In this case, both will apply with certainty. Equating the expected value of the item associated with $x = 0$ with that obtained for each possible offer we obtain:

$$1000 - 300 = 10,000B(x) + 1000(1 - B(x)) - x - 300, \quad 0 \leq x \leq a.$$

Thus B is a uniform distribution on $[0, 9000]$.

Suppose next that there are three potential applicants. In equilibrium it cannot happen that all three will participate with certainty, since then an offer of 0 clearly entails a loss of the application cost. Thus there exists a probability, say p , that a potential applicant will submit an application. This means that the expected net value is identical for an individual who applies and another who does not, and its size must be 0. The probability that an individual applies with an offer greater than x is $p(1 - B(x))$. The item of \$10,000 will be obtained by an applicant who offers x if the other two do not offer more. The other item will be obtained if exactly one of them offers more. Thus we have for all x in $[0, a]$:

$$0 = [1 - p(1 - B(x))]^2 10,000 + 2[1 - p(1 - B(x))]p(1 - B(x))1000 - x - 300.$$

Now, $p \approx 0.9$ is computed by letting $B(0) = 0$, and $a = 9700$ is computed by letting $B(a) = 1$.

3. Queues with Payments: Identical Customers

We consider a single server queueing model. This is done to simplify the discussion in §5. We assume that:

(i) The potential demand for service consists of a stationary stream of identical risk neutral customers with average arrival rate of λ_0 per unit time.

(ii) The service requirements of individual customers are exponentially, independently, and identically distributed with mean $1/\mu$.

(iii) The service value to a customer is R .

(iv) The customer's time cost is c per unit time. We assume $R \geq c/\mu$, otherwise no customer will be ready to arrive. We do not explicitly consider a fixed cost associated with joining the queue because every arriving customer is eventually served, so that if such a cost exists it can be subtracted from the service value.

(v) At the time a customer's need for service arises, he does not know the queue size, but he is well informed about its statistical distribution, on which he is basing his decision whether to join the queue or not.

(vi) Social welfare consists of the average value of service minus waiting costs per unit time.

For completeness, we review the first-come first-served model. Consider a facility with an admission toll θ . Denote by $\bar{w}(\lambda)$ the expected time a customer spends in the system when the arrival rate of customers joining the queue is λ . If every potential customer joins the queue then each customer's net benefit is $R - \theta - c\bar{w}(\lambda_0)$. If this value is negative then the equilibrium arrival rate, λ , must satisfy

$$\theta + c\bar{w}(\lambda) = R. \tag{3.1}$$

If this value is positive then the arrival rate will be $\lambda = \lambda_0$.

Consider the profit maximizer's problem first. The profit-maximizing admission toll never yields a positive customer surplus, since in such a case it can be increased without reducing the arrival rate. Therefore, (3.1) holds under the profit-maximizing toll. The server's problem is then to choose an admission toll θ in order to maximize $\theta \cdot \lambda$ subject to

$$0 \leq \lambda \leq \lambda_0, \quad \text{and} \quad R - \theta - c\bar{w}(\lambda) = 0. \tag{3.2}$$

The social objective is to maximize $\lambda(R - c\bar{w}(\lambda))$ subject to $0 \leq \lambda \leq \lambda_0$. By (3.1) this is exactly the profit maximizer's objective. We conclude, therefore, that the profit maximizer's objective is identical to the social one, as was shown by Edelson and Hildebrand (1975).

We now add the following assumptions:

(vii) When entering the queue the customer chooses a nonnegative amount to pay the server. No customer knows the actual amounts paid by others, but each knows their statistical distribution.

(viii) A customer is placed in the queue ahead of all those who paid smaller amounts. This may mean that the service of a customer is interrupted.

(ix) When a customer whose service has been interrupted returns to service, the service is resumed from the point where it has been stopped with no loss of service.

(x) Customers' payments are considered transfer of income that do not affect social welfare.

The equilibrium behavior of customers is characterized by a probability that a potential customer joins the queue (or equivalently, an arrival rate of customers) and a distribution of payments among those who join. Equilibrium requires that all customers have identical expected welfare. This value is clearly zero for a customer who decides not to arrive. It is equal to the service value minus the payment minus the expected waiting cost given the payment for a customer who joins. Let B denote an equilibrium cumulative probability distribution of payments for arriving customers. Considerations identical to those of §2 can be used to prove that (2.1) and (2.2) are valid also in this case. We will now prove (2.3). Let $w_x(\lambda)$ denote the expected time a customer who paid x spends in the system. We refer to $w_x(\lambda)$ as the *waiting time* of the customer. Note that it includes both his wait in the queue and his service time. Then (2.1) and (2.2) imply that for some constant a ,

$$x + cw_x(\lambda) \text{ is constant for } 0 \leq x \leq a. \quad (3.3)$$

In particular,

$$x + cw_x(\lambda) = cw_0(\lambda) \text{ for } 0 \leq x \leq a. \quad (3.4)$$

A customer who pays the maximum amount a is guaranteed to be served immediately without interruption, so that his waiting time is $w_a(\lambda) = 1/\mu$. Substituting in (3.4) we obtain

$$a = c(w_0(\lambda) - 1/\mu). \quad (3.5)$$

By (ii), because of the memoryless characteristic of the exponential distribution, the residual service of each customer in the system has the same (exponential) distribution. Therefore, the order in which customers are

served is unimportant from the social point of view. This implies that the effect on social welfare caused by a customer who arrives is independent of his payment.

In equilibrium an arriving customer is indifferent among all possible payments in $[0, a]$. In particular, customers join as long as it is worth doing so with no payment at all (i.e., with $x = 0$). By (2.1) and (2.2), a customer who offers no payment is certain to be placed at the end of the queue and to stay there until his service is completed, imposing no extra wait on the others. Since the total wait in the system is independent of the order of the service, we conclude that a customer joins if and only if he finds it worthwhile doing so when he bears all the additional waiting costs resulting from his arrival. But this is exactly the social criterion! Thus the considerations of the individual and the social planner are identical, as claimed.

We now provide an alternative explanation for this result. As explained above, the social cost imposed by a customer who joins the queue is $cw_0(\lambda)$, independent of his payment. A customer who pays x waits only $w_x(\lambda)$, and the costs he incurs contribute $cw_x(\lambda)$ to the social costs. Therefore, the difference $c[w_0(\lambda) - w_x(\lambda)]$ expresses the externalities he imposes on others. However, by (3.4) this expression equals x , so that in equilibrium *the amount paid by a customer equals the externalities he causes*. This again explains why the individual's behavior is socially optimal. (It is interesting to compare this outcome with that of Dewan and Mendelson (1990), who show, in a general queueing model, that a fixed price of size equal to the expected externality can be imposed in order to optimally regulate the arrival rate.)

Suppose now that Assumptions (ii) and (ix) are dropped, but we still assume that the service demands are independently and identically distributed. The individual's decision consists of two parts: Whether to join the queue, and if so, also how much to pay. We first note that although *for a given state of the queue and a given distribution of payments* there may be social preference with regard to a payment of a joining customer (in order to avoid preemptions or serve a customer with a short residual service first), the unconditional distribution of payments is irrelevant to social welfare. All that matters is that payments are used to rank the joining customers in a way that is independent of their arrival

times. Thus the only part in the customer's decision that matters is the probability that he joins.

As before, customers will join as long as it is worth doing so without any payment, and this fact is independent of the service distribution. Again, in this case the customer imposes no externalities and, therefore, the rate by which customers join will be socially optimal. Note that although the rate of arrivals is optimally regulated, the discipline is not socially optimal due to the losses caused by preemptions and the excess wait caused by not serving customers according to their residual service times. A possible correction is to exclude preemptions, though the resulting discipline still will not be socially optimal because a customer with zero payment now imposes some externalities since his service cannot be preempted.

4. Social Welfare and Server's Revenue

Social welfare is

$$SW = \lambda(R - c\bar{w}(\lambda)), \quad (4.1)$$

where the expected waiting time, $\bar{w}(\lambda)$, is independent of the order of service. This last observation follows from the assumption of an exponential service with no service loss as a result of preemption. Thus, $\bar{w}(\lambda)$ is identical to the expected wait in a similar queue with an average arrival rate of λ and a first-come, first-served discipline.

The server's revenue consists of the payments obtained from those customers who join the queue:

$$\Pi = \lambda \int_0^a x dB(x), \quad (4.2)$$

where a is defined in (3.5). Substituting x from (3.4) we obtain

$$\begin{aligned} \Pi &= \lambda \int_0^a c[w_0(\lambda) - w_x(\lambda)] dB(x) \\ &= \lambda c[w_0(\lambda) - \bar{w}(\lambda)]. \end{aligned} \quad (4.3)$$

Comparing (4.1) with (4.3) we find that

$$SW = \Pi + \lambda[R - cw_0(\lambda)], \quad (4.4)$$

where the term $R - cw_0(\lambda)$ is the customer's surplus after deduction of payments. If not all of the potential

demand is served, customers are indifferent between joining the queue and not, so that this value is zero and $SW = \Pi$.

5. Optimal Rate of Service

In this section we assume that the service rate μ is a decision variable, and that the cost of operating a facility with a service rate of μ is $g(\mu)$ per unit time. As shown in Section 3, in a first-come, first-served system with an admission toll the social and profit-maximizing objectives are identical. Consequently, the service rate chosen by a profit-maximizing server is socially optimal. When customer payments are introduced we showed that for any given service rate the equilibrium arrival rate is socially optimal and, therefore, the socially optimal service rate is as in the first-come, first-served case. We now compare this rate to the one chosen by a profit-maximizing server.

If the server chooses a slow rate, then in equilibrium a portion of the potential demand will choose not to join the queue. If, on the other hand, the server supplies very fast service then all the potential demand (with rate λ_0) will be served. Therefore, there exists a cut-off value, μ_0 , such that $\lambda = \lambda_0$ if and only if the service rate is at least μ_0 .

We consider first the case where $\mu < \mu_0$. An increase in μ has two contradicting effects on the average wait: each customer's service is made shorter, but more customers are attracted to join the queue. However, since customer behavior is socially optimal, social welfare increases when the service is made faster (it would increase even if λ remains unchanged, all the more so when it changes to its new optimal value). In this range $\Pi = SW$, since customers are indifferent between joining the queue and balking, and all social welfare goes to the server. Therefore,

$$\frac{d\Pi}{d\mu} = \frac{dSW}{d\mu} > 0, \quad \text{for } \mu < \mu_0. \quad (5.1)$$

We now consider the case where $\mu > \mu_0$. An increase in μ will not affect the arrival rate, which already consists of the whole potential demand. Thus, the average wait must decrease, increasing social welfare. By (4.3) the server's revenue decreases, because when the service is made faster the difference between the expected wait

of a low-priority customer and the expected wait of an arbitrary customer is made smaller. Thus the same number of customers is served, but customers pay less. We conclude therefore that

$$\frac{d\Pi}{d\mu} < 0, \quad \frac{dSW}{d\mu} > 0, \quad \text{for } \mu > \mu_0. \quad (5.2)$$

When μ increases to infinity then the expected waiting time, even of the customer with the lowest priority, as well as the expected service time, decrease to zero. Therefore,

$$\lim_{\mu \rightarrow \infty} SW = \lambda_0 R, \quad \lim_{\mu \rightarrow \infty} \Pi = 0. \quad (5.3)$$

Figure 1 illustrates the conclusions obtained above. We see that if the service rate can be controlled without cost then μ_0 is the rate chosen by the profit maximizer; all the potential demand arrives to the queue, and customer's welfare is zero. Clearly the socially optimal rate is infinite.

Recall that operating a facility with a service rate of μ involves a cost of $g(\mu)$ per unit time. The following two cases are possible:

(i) The socially optimal rate is larger than μ_0 , as for g_1 in Figure 1, where the optimal rate is μ_1 . In this case the server will choose a slower service speed (frequently this will be exactly μ_0 , as in Figure 1).

(ii) The socially optimal rate is smaller than or equal to μ_0 . This possibility is shown for the cost function g_2

in Figure 1. In this case the server will voluntarily choose this optimal rate (μ_2 in Figure 1), because it also maximizes his profits.

6. Heterogeneous Population of Potential Customers

In this section we consider customers with different service valuations, waiting costs, and service requirements.

6.1. Different Service Valuations

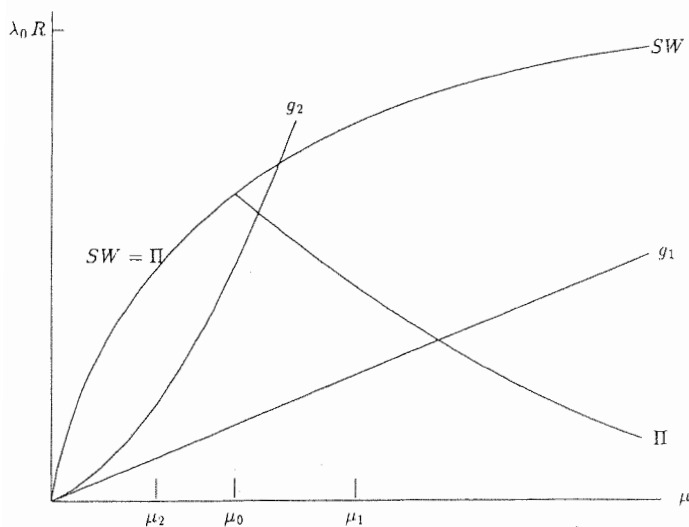
Suppose that potential customers value the good differently. It is (socially) desired that service be rendered to customers whose service valuation is higher. We first show that this property is attained in equilibrium. Let $w(x)$ be the equilibrium expected waiting time for a customer whose payment is x . Clearly, $w(x)$ is monotone decreasing. A potential customer who values the service at R faces the following problem:

$$\max \{0, \max_{x \geq 0} \{R - x - cw(x)\}\}. \quad (6.1)$$

If the maximum of the inside expression is positive he will arrive with the maximizing x -value, if it is negative he will not arrive, and if it is zero he is indifferent between these options. It is easily seen that individuals who decide to arrive have valuations not smaller than those who do not apply. Note, however, that once a customer decides to arrive, his next decision with regard to the size of his payment is independent of his service valuation. The equilibrium payment distribution for the arriving customers is computed in exactly the same way as if they come from a homogeneous population, and in particular (2.1) and (2.2) hold.

Denote by R^* the lowest valuation of an individual who arrives in equilibrium. Such a customer offers zero payment. Consider an individual who does not arrive, thus having $R \leq R^*$. If he is persuaded to arrive then his contribution to social welfare is $R - cw(0)$, no matter what payment he chooses. But this individual, while deciding not to arrive, already considered the possibility of arriving with no payment and rejected it. Hence $R - cw(0) \leq 0$ and $R - cw(0) \leq 0$, and the change does not increase social welfare. Similarly, we argue that it cannot happen that social welfare increases by persuading customers not to arrive. If we do so, then the best choice is a customer with valuation R^* . But such a customer is ready to arrive with no payment, in which

Figure 1



case he imposes no externalities. Thus, the change in social welfare entailed by persuading him not to arrive is nonpositive. We conclude that the optimal arrival rate is attained.

6.2. Different Service Requirements

Suppose that in addition to different service valuations, customers differ by their service requirements as expressed by different service rates μ . A customer with parameters R and μ faces the following problem:

$$\max \{0, \max_{x \geq 0} \{R - x - c(w_q(x) + 1/\mu)\}\}, \quad (6.2)$$

where $w_q(x)$ is the expected waiting time in the queue while others are served. Let $R' = R - c/\mu$, then (6.2) is similar to (6.1) and again, the payments of those who decide to arrive are independent of their parameters. In particular, the order in which they are served will be independent of their service requirements. This is not compatible with social optimality, which requires serving customers in increasing order of their service rates. A similar problem exists with Mendelson and Whang's (1990) priority prices, and to overcome it they suggested computing prices that are based on the *actual* service requirements. Similar ideas can be used in our model to induce higher payments by customers with higher service rates, but the resulting mechanism will not be self-regulating any longer.

6.3. Different Waiting Costs

Suppose now that customers have identical service rates, but they differ by their service valuations and waiting costs. We claim that also in this case the socially optimal behavior is induced. We first observe that once a customer decides to arrive, his payment is independent of his service valuation. Moreover, it is clear that the individual optimization results in higher payments and priorities for customers with higher waiting costs, among those who decide to arrive. It is less obvious that the division between arrivals and non-arrivals conforms with social optimality. To demonstrate this property let us assume for simplicity that there is just a finite set $c_1 < c_2 < \dots < c_n$ of possible waiting costs. A customer with waiting cost of c_i per unit time is called an i -customer. All the i -customers who decide to arrive will have in equilibrium the same expected waiting times

(though not the same expected welfare since they may differ by their service valuation). From considerations similar to those in §2, the cumulative probability distribution of i -customers will be continuous and strictly increasing on an interval $[a_{i-1}, a_i]$ with $a_{i-1} < a_i$ and $a_0 = 0$. The 1-customers impose externalities on other 1-customers only. As in §2, such a customer's payment is equal to the additional wait he imposes on others, and a 1-customer arrives if and only if it is socially desirable that he arrive.

Suppose inductively that this property holds for $(i-1)$ -customers. An i -customer who pays a_{i-1} causes the same external effects as an $(i-1)$ -customer who offers this amount. (This is where we use the assumption that the service rates are identical.) If he chooses to increase his payment (within the interval $[a_{i-1}, a_i]$), he will reduce his expected waiting costs but not his expected welfare (since the latter is the same for all i -customers). Thus, the extra payment is equal to the expected saving in waiting costs, which in turn is equal to the expected additional waiting costs to other i -customers. Thus, we conclude that if he pays $a_{i-1} + g$, then by the inductive assumption, a_{i-1} is the part of the externalities imposed on k -customers, $k < i$, and g is equal to the part imposed on i -customers. Altogether, the payment is again equal to the externalities, and a customer arrives only if the sum of his waiting cost and the externalities he imposes is less than his service valuation. Thus the individual optimization in equilibrium is again identical to social optimization.

6.4. The Optimal Rate of Service

We will show now that the qualitative result of §5, namely that the profit-maximizing rate of service is smaller than or equal to the socially optimal rate, is not restricted to the case of homogeneous customers. To simplify the presentation we will assume two customer types. Furthermore, we assume that the arrival process is Poisson and use explicit formulas that are available for this model.

We first claim that it is sufficient to prove that the total expected rate of customer surplus, that is $SW - \Pi$, is a nondecreasing function of μ . Recall that SW and Π denote social welfare and profits from customer payments, respectively. To prove this claim, let μ_s be the welfare-maximizing rate. Let $g(\mu)$ be the cost per time

unit involved in providing service rate of μ . Then for $\mu > \mu_s$, by definition of μ_s

$$SW(\mu_s) - g(\mu_s) \geq SW(\mu) - g(\mu),$$

and since $SW - \Pi$ is nondecreasing

$$SW(\mu_s) - \Pi(\mu_s) \leq SW(\mu) - \Pi(\mu).$$

Hence,

$$\begin{aligned} \Pi(\mu_s) - g(\mu_s) &= (SW(\mu_s) - g(\mu_s)) - (SW(\mu_s) - \Pi(\mu_s)) \\ &\geq (SW(\mu) - g(\mu)) - (SW(\mu) - \Pi(\mu)) \\ &= \Pi(\mu) - g(\mu). \end{aligned}$$

Thus, the profit maximizer prefers μ_s to any $\mu > \mu_s$.

Assume that there are two classes of customers, with potential demand rates $\lambda_{1,0}$ and $\lambda_{2,0}$, service evaluations R_1 and R_2 , and unit time costs c_1 and c_2 , respectively. Assume, without loss of generality, that $c_1 < c_2$. If the equilibrium arrival rates λ_1 and λ_2 are both positive, then for some constants $0 < a < b$, type 1 customers will offer payments in $[0, a]$ and type 2 customers will offer payments in $[a, b]$. Let $\hat{w}(\lambda)$ be the expected waiting time of a customer who sees an arrival rate λ of higher priority customers. Then,

$$\hat{w}(\lambda) = \frac{1}{\mu(1 - \lambda/\mu)^2}.$$

[This equation can be found in Kleinrock (1967), or it can be derived from the fact that $\hat{w}(\lambda)$ is equal to $(L + 1)B$, where $L = \lambda/(\mu - \lambda)$ is the expected number of customers in the system at the time of arrival, and $B = 1/(\mu - \lambda)$ is the expected length of a busy period in an $M/M/1$ queue with parameters λ and μ .]

Let S_i denote the equilibrium expected surplus of customers of type i . Let $S = \lambda_1 S_1 + \lambda_2 S_2$ denote the total rate of customers surplus. Then, $S = SW - \Pi$, and we want to prove that S is nondecreasing in μ . Since S is a continuous function of μ it is sufficient to prove the monotonicity claim in each of the cases listed below.

If $\lambda_1 > 0$ then

$$S_1 = R_1 - c_1 \hat{w}(\lambda_1 + \lambda_2) = R_1 - (a + c_1 \hat{w}(\lambda_2)), \quad (6.3)$$

and if $\lambda_2 > 0$ then

$$S_2 = R_2 - (a + c_2 \hat{w}(\lambda_2)). \quad (6.4)$$

For small values of μ , we will have in equilibrium $\lambda_1 < \lambda_{1,0}$ and $\lambda_2 < \lambda_{2,0}$. In this case both types of customers have zero surplus and $S = SW - \Pi = 0$.

For higher values of μ we obtain one of the following cases:

(i) $\lambda_1 = \lambda_{1,0}$ and $\lambda_2 = 0$ or $\lambda_1 = 0$ and $\lambda_2 = \lambda_{2,0}$. This case is similar to that of a homogeneous population, and monotonicity of S is proved as in §5.

(ii) $0 < \lambda_1 < \lambda_{1,0}$ and $\lambda_2 = \lambda_{2,0}$. In this case we can substitute $S_1 = 0$ in (6.3) and with (6.4) obtain

$$S_2 = (R_1 - R_2) - (c_2 - c_1) \hat{w}(\lambda_{2,0}).$$

Thus,

$$\frac{dS}{d\mu} = \lambda_{2,0} \frac{dS_2}{d\mu} = -\lambda_{2,0}(c_2 - c_1) \frac{d\hat{w}(\lambda_{2,0})}{d\mu} > 0.$$

(iii) $\lambda_1 = \lambda_{1,0}$ and $0 < \lambda_2 < \lambda_{2,0}$. In this case we can substitute $S_2 = 0$ in (6.4) and with the second equality in (6.3) obtain

$$c_1 \hat{w}(\lambda_{1,0} + \lambda_2) = R_2 - (c_1 - c_2) \hat{w}(\lambda_2).$$

Differentiating with respect to μ we obtain

$$\begin{aligned} 2 \frac{d\lambda_2}{d\mu} &\left(\frac{c_1}{\left(1 - \frac{\lambda_{1,0} + \lambda_2}{\mu}\right)^3} + \frac{(c_2 - c_1)}{\left(1 - \frac{\lambda_2}{\mu}\right)^3} \right) \\ &= \frac{c_1 \left(1 + \frac{\lambda_{1,0} + \lambda_2}{\mu}\right)}{\left(1 - \frac{\lambda_{1,0} + \lambda_2}{\mu}\right)^3} + \frac{(c_2 - c_1) \left(1 + \frac{\lambda_2}{\mu}\right)}{\left(1 - \frac{\lambda_2}{\mu}\right)^3}. \end{aligned}$$

Since $\lambda_{1,0} > 0$, it follows that

$$2 \frac{d\lambda_2}{d\mu} < 1 + \frac{\lambda_{1,0} + \lambda_2}{\mu}.$$

On the other hand, by (6.3)

$$S_1 = R_1 - \frac{c_1}{\mu \left(1 - \frac{\lambda_{1,0} + \lambda_2}{\mu}\right)^2}.$$

Differentiating with respect to μ we obtain that $dS_1/d\mu$ is proportional to $1 + (\lambda_{1,0} + \lambda_2)/\mu - 2 d\lambda_2/d\mu$. By the previous inequality this derivative is positive.

(iv) $\lambda_1 = \lambda_{1,0}$ and $\lambda_2 = \lambda_{2,0}$. S increases with μ since the expected waiting time decreases while the same service is given.

7. Concluding Remarks

Section 2 discussed in general the effects of payments offered by identical applicants on the application process. We have shown that a mechanism which clearly may be useful to discriminate among individuals of different characteristics can also be used to motivate identical individuals to behave in a socially desired manner. A more detailed analysis was possible for the queueing mode. We demonstrated that the price mechanisms suggested by Edelson and Hildebrand (1975) for a homogeneous population and by Mendelson and Whang (1990) for a heterogeneous population can be replaced by a self-regulating mechanism. Specifically, customers are free to choose any nonnegative payment, instead of selecting a payment from a finite set specified by the system organizer. Note that there are also intermediate possibilities where the payments are restricted to certain domains instead of to a discrete set or to the nonnegative values. Mendelson and Whang's (1990) priority prices were carefully computed so that the equilibrium payment will exactly split the population by their characteristics, so that in each probability mass the customers are identical. However, these prices are not unique. For example, in the case of identical customers, if the payments are restricted to be in $[0, a]$, for a value of a smaller than that of equation (3.5), then the optimal arrival rate will still be obtained. The payment distribution will be continuous on $[0, a)$ and will have a probability mass at a . However, since the customers are identical, the order in which those who pay a will be served is unimportant. All that is needed is that the allowed domain should include an interval $[0, a]$ for some $a > 0$.

A related problem is the optimal design of a contest (see Glazer and Hassin 1987). In a contest participants, possibly with different abilities, independently decide on the amount of effort to invest in preparing for the contest. The organizer of the contest has a budget to be allocated as prizes that will be awarded according to the participants' outcomes, the higher prizes given to the participants with the higher outcome. Contests are similar to the model analyzed here but with some important differences: The "bids" are actual costs and not transfer payments, and the prizes are decision variables designed to maximize total output.

Another conclusion of this paper is that when the

speed of service is a decision variable, a profit maximizer may choose too slow a rate relative to the socially optimal one. The service rate can be controlled in several ways, and each case may require different analysis. The simplest way, which is the one implicitly assumed here, is controlling the work process thereby affecting the service time of each individual customer. Another possibility is controlling the number of servers. In this case μ is restricted to integral multiples of the service rate of a single server. The analysis is only slightly changed in this case, and similar results with obvious modifications are obtained.

Another interesting case is one in which the facility is not continuously open but rather accepts customers during specific time intervals. In this case, the definition of the service rate which is relevant for our discussion differs from the common one. For example, suppose that demand is generated with rate λ , that μ customers can be served per unit time, and that the facility is open only half of the day. The effective rate of demand is then 2λ , or equivalently the effective service rate is $\mu/2$. The server can control the service "speed" by controlling the size of the interval during which the facility is open. The analysis of this case is significantly different since customers who wish to be served must decide not only on the payment they offer but also on their arrival time. Some will even arrive before the opening time of the facility to assure themselves a favorable position in the queue. The equilibrium arrival pattern in this case typically will be time-dependent, as derived for the first-come, first-served discipline by Glazer and Hassin (1983).¹

A paper by Holt and Sherman (1982) considers the allocation of a finite number of prizes at a specified time on a first-come, first-served basis to individuals who independently choose their arrival times. We note that such systems are inefficient from the social point of view. A simple alternative eliminates the social costs asso-

¹ The transition equations in the paper by Glazer and Hassin (1983) are valid only under the assumption that the total number of arrivals during a day is Poisson. This is in contrast to the more general assumption made there. However, this assumption is a natural one, obtained when each person in a large population independently decides whether to arrive. Without this assumption, a more involved set of transition equations is necessary, taking into account the dependence of future rate on the current state.

ciated with the first-come, first-served rule, namely allocation of the prizes on a random basis to those present at the time of distribution. This makes early arrivals worthless. However, if it is costly to show up at the distribution point, it may be socially desired that some of the potential demanders should not show up at all. In this case again, individual optimization will create excess congestion. Bidding, as suggested in the present paper, eliminates the unnecessary wait and induces the right rate of show ups.

References

- Dewan, S. and H. Mendelson, "User Delay Costs and Internal Pricing for a Service Facility," *Management Sci.*, 36 (1990), 1502-1517.
- Dolan, R. J., "Incentive Mechanisms for Priority Queueing Problems," *Bell J. Economics*, 9 (1978), 421-436.
- Edelson, N. M. and K. Hildebrand, "Congestion Tolls for Poisson Queueing Processes," *Econometrica*, 43 (1975), 81-92.
- Glazer, A. and R. Hassin, "?/M/1: On the Equilibrium Distribution of Customer Arrivals," *European J. Operational Res.*, 13 (1983), 146-150.
- and —, "Stable Priority Purchasing in Queues," *Oper. Res. Letters*, 4 (1986), 285-288.
- and —, "Optimal Contests," *Economic Inquiry*, 26 (1988), 133-143.
- Grassmann, W. K., "The Economic Service Rate," *J. Operational Res. Society*, 30 (1979), 149-155.
- Hassin, R., "On the Optimality of First-come Last-served Queues," *Econometrica*, 53 (1985), 201-202.
- , "Consumer Information in Markets with Random Product Quality: The Case of Queues and Balking," *Econometrica*, 54 (1986), 1185-1195.
- Haviv, M., "Stable Strategies for Processor Sharing Systems," *European J. Operational Res.*, 52 (1991), 103-106.
- Holt, C. A., Jr. and R. Sherman, "Waiting-line Auctions," *J. Political Economy*, 90 (1982), 280-294.
- Kleinrock, L., "Optimal Bribing for Queue Position," *Oper. Res.*, 15 (1967), 304-318.
- Lui, F. T., "An Equilibrium Model of Bribery," *J. Political Economy*, 93 (1985), 760-781.
- Marchand, M. G., "Priority Pricing," *Management Sci.*, 20 (1974), 1131-1140.
- Mendelson, H. and S. Whang, "Optimal Incentive-compatible Priority Pricing for the M/M/1 Queue," *Oper. Res.*, 38 (1990), 870-883.
- Mills, D. E., "Ownership Arrangements and Congestion-prone Facilities," *The American Economic Review*, 71 (1981), 493-502.
- Myrdal, G., *Asian Drama: An Inquiry into the Poverty of Nations*, Pantheon, NY, 1968.
- Naor, P., "The Regulation of Queue Size by Levying Tolls," *Econometrica*, 31 (1969), 15-24.
- Samuelson, W. F., "Competitive Bidding with Entry Costs," *Economic Letters*, 17 (1985), 53-57.
- Stidham, S., Jr., "Optimal Control of Admissions to a Queueing System," *IEEE Transactions on Automated Control*, AC-30 (1985), 705-713.

Accepted by Linda Green; received February 18, 1992. This paper has been with the author 3 months for 1 revision.