

Congestion Tolls for Poisson Queuing Processes

Author(s): Noel M. Edelson and David K. Hilderbrand

Source: *Econometrica*, Jan., 1975, Vol. 43, No. 1 (Jan., 1975), pp. 81-92

Published by: The Econometric Society

Stable URL: <https://www.jstor.org/stable/1913415>

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at <https://about.jstor.org/terms>



The Econometric Society is collaborating with JSTOR to digitize, preserve and extend access to *Econometrica*

JSTOR

CONGESTION TOLLS FOR POISSON QUEUING PROCESSES

BY NOEL M. EDELSON AND DAVID K. HILDEBRAND

The relationship between Pareto optimal (θ_p) and revenue maximizing (θ_r) tolls is examined for queuing models that permit balking. When customers have the same value for waiting time, $\theta_s \equiv \theta_r$, provided the entrepreneur can impose a simple two-part tariff. With heterogeneous values for waiting time, θ_r can be greater than, equal to, or less than θ_s . Expanding the number of servers and charging multi-part tariffs are shown to be alternative methods for segmenting the market, and the welfare implications of these two strategies are explored.

1. REVIEW OF NAOR'S MODEL

P. NAOR [3] has examined the relationship between Pareto optimal and revenue maximizing tolls for a queuing model that permits balking. Consider an $M/M/1$ queue with gross accession rate λ and service rate μ . Because an arriving customer need not join the queue, existence of a steady state does not require that $\rho = \lambda/\mu$ be less than unity. Each customer has a cost per unit of service and waiting time, c , and receives a benefit R if he is served by the facility. If an arriving customer finds q people ahead of him, he faces an expected waiting plus service time of $(q + 1)/\mu$. The toll charged by the facility, θ , determines a critical queue size, n , such that the customer balks if $q \geq n$. Assuming that there is no specific balking cost, the customer's decision rule is: join queue if $\theta + c(q + 1)/\mu \leq R$ (if $q < n$); balk if $\theta + c(q + 1)/\mu > R$ ($q \geq n$).

Expressions (i), (ii), and (iii) report Naor's results for expected queue size, $E[q]$, the expected number of customers diverted per unit time, ζ , and the expected number of customers joining the queue per unit time, $\lambda - \zeta$.

$$(i) \quad E[q] = \frac{\rho}{1 - \rho} - \frac{(n + 1)\rho^{n+1}}{1 - \rho^{n+1}},$$

$$(ii) \quad \zeta = \frac{\lambda\rho^n(1 - \rho)}{1 - \rho^{n+1}},$$

$$(iii) \quad \lambda - \zeta = \lambda \frac{(1 - \rho^n)}{1 - \rho^{n+1}}.$$

Naor's social welfare function, P , is expected benefits per unit time, $P = (\lambda - \zeta)R - cE[q]$. Expected revenue per unit time, M , is given by $M = (\lambda - \zeta)\theta = (\lambda - \zeta)(R - cn/\mu)$.

The point of Naor's article is to show that the value of n which maximizes P is greater than that which maximizes M , i.e., the revenue maximizing toll exceeds the

socially optimal toll.¹ Yechiali [5] reaches a similar conclusion if a penalty charge is imposed for balking. These results differ from that obtained for a non-stochastic model by Edelson [1]; there it is proved that when customers have the same value for time, the Pareto optimal and revenue maximizing tolls are identical.

In the next section we demonstrate that the same toll maximizes P and M in an $M/M/1$ queue where no balking is allowed, and explain why balking upsets that equivalence. It is also shown that, even if balking is allowed, the revenue maximizing and Pareto optimal tolls are identical if the server sells in advance rights to service, with a predetermined toll if service is rendered.

2. EQUIVALENCE OF θ_s AND θ_r WHEN ALL CUSTOMERS HAVE THE SAME VALUE FOR TIME

We can interpret Naor's R , the benefit from service, as the total cost of being served at an alternate facility. Let the money toll at that facility be τ , which we suppose equals the real resource cost per customer; expected waiting plus service time there, γ , is assumed to be independent of customer flow to avoid reciprocal externalities. Following Naor we assume Poisson arrival and service distributions with parameters λ and μ respectively, but customers are not allowed to balk. An irrevocable decision to join one of the queues must be made before observing the state of the system.

Although a customer does not know in advance his actual waiting plus service time, he is assumed to have an expectation, based on past experience, of what the expected time cost will be at each facility. Suppose that a customer patronizes one facility exclusively, and the set of potential customers is partitioned such that a fraction $\lambda(\theta)/\lambda$ patronize the θ -facility and a fraction $1 - \lambda(\theta)/\lambda$ the τ -facility, where λ is the exogenous total arrival rate. This partition ensures that the arrival rate at the θ -facility is $\lambda(\theta)$ and $\lambda - \lambda(\theta)$ at the τ -facility. The system is said to be in statistical equilibrium if the partition is one where no customer has an incentive to change servers because his subjective estimates of expected waiting plus service time differ from those determined by arrival rates $\lambda(\theta)$ and $\lambda - \lambda(\theta)$.

Therefore, if θ is to be an equilibrium toll, $\lambda(\theta)$ must be such that customers are indifferent (ex-ante) between patronizing the two facilities. Equating total expected costs, we have

$$(1) \quad \theta + cE[t; \theta] = \tau + c\gamma,$$

where $E[t; \theta]$ is expected waiting plus service time given a toll θ . As Little [2] has shown for general queuing processes, in a steady state $E[t; \theta]$ equals expected

¹ A peculiar feature of this model, apparently overlooked by Naor, is that a revenue maximizer may find it profitable to expand resources for reducing $1/\mu$, when this is inappropriate from a social point of view. In a deterministic model, a downward shift of the marginal cost curve must increase the sum of consumer plus producer surplus for a competitive industry by more than it increases the profits of the same industry operated by a monopolist. But with Naor's model, if μ rises from 1 to $5/2$ while $R = 9$, $c = 2$, $\lambda = 0.5$, $\Delta M \approx 1.39$ (optimal n rises from 1 to 2) and $\Delta P \approx 1.27$ (optimal n is unchanged at 3). Therefore, if the cost per time of $\Delta\mu = 5/2 - 1$ were between 1.39 and 1.27, a monopolist would invest in increasing the service rate although expected social benefits are lowered by this action.

queue size, $E[q; \mathbf{4}]$, divided by the arrival (departure) rate $\lambda(\theta)$:

$$(2) \quad E[t; \theta] = \frac{E[q; \theta]}{\lambda(\theta)} = \frac{\rho(\theta)}{\lambda(\theta)[1 - \rho(\theta)]} = \frac{1}{\mu - \lambda(\theta)}.$$

If $\theta \cong \tau$ and $E[t; \theta] \cong \gamma$, facilities charging a higher toll must offer shorter expected waiting plus service times. A higher toll reduces expected waiting time by decreasing the arrival rate, i.e., $\lambda'(\theta) < 0$. Of course, a customer who arrives when the actual queue size exceeds (is less than) \tilde{q} will be disappointed (pleased) that he chose the facility charging θ , where

$$\tilde{q} = \frac{\rho(\theta)}{1 - \rho(\theta)}.$$

Now consider the values for θ that maximize expected social welfare and expected revenue. Expected social welfare per unit time equals expected gross benefits per unit time (gross benefits per customer times expected customers per unit time) less expected service costs per unit time (c times expected queue size). Toll revenue is not included in the objective function, since it is simply a transfer of income from one social group (customers) to another (owners of the facility). The socially optimal toll, θ_s , must be such that $\lambda(\theta)$ maximizes

$$(3) \quad \lambda(\theta)(\tau + c\gamma) - cE[q; \theta],$$

where $\lambda(\theta)$ and $E[q; \theta]$ are determined implicitly by equation (1).

A revenue maximizer, on the other hand, will seek to maximize $\lambda(\theta) \cdot \theta$. Let the revenue maximizing toll be θ_r . By (1), θ_r equals expected net benefits per customer.

$$(4) \quad \theta_r = \tau + c\gamma - cE[t; \theta_r] = \tau + c\gamma - \frac{cE[q; \theta_r]}{\lambda(\theta_r)}.$$

Therefore, the entrepreneur's objective function is identical to (3), and $\theta_s \equiv \theta_r$.

The objective functions are different in Naor's model because balking makes expected net benefits per customer greater than θ ; if arrivals do not join when $q \geq n$, expected queue size must be less than n . Those customers arriving when $q < n$ obtain inframarginal benefits which are included in P but not in M . A private entrepreneur counts only the toll he receives, and by making $\theta_r > \theta_s$ he is able to expropriate some of his customers' inframarginal benefits. Because social welfare is below its maximal value, the extra revenue attained at θ_r is less than the additional cost imposed on balking customers.

The no-balking model $M/M/1$ is easily extended to incorporate heterogeneous values for c . The structure of the model and its qualitative results are identical to those obtained by Edelson [1] for a non-stochastic problem: (1) a Pareto optimal toll must be strictly positive and not equal to τ , i.e., $\theta_s > 0$ and $\theta_s < \tau$ or $\theta_s > \tau$; (2) if $\theta_s > \tau$, the revenue maximizing toll *may* be below θ_s , i.e., a private entrepreneur may attract a customer flow larger than the Pareto optimal rate. A sketch of the $M/M/1$ model, which assumes a distribution of values for c over the population of customers, is included in Appendix A.

It can also be shown that $\theta_s \equiv \theta_r$ if the entrepreneur can impose a two-part tariff, selling rights to service valid for a predetermined period with a specific toll if service is rendered.² The length of the validation period is independent of the actual queue size when the right is issued, and customers who are waiting or in service when their rights expire are served. It is also assumed that the validation period is sufficiently long so that we can ignore transient effects before the queuing process reaches a new statistical equilibrium following, e.g., a change in θ . To preserve the constancy of c we assume either that a customer demands at most one service during the validation period or that the value of each service is independent of the number of requests. The equilibrium price for such a right is the customer's expected gain,

$$(5) \quad \hat{\lambda} \sum_{i=0}^{n-1} \pi_i \left[\tau + c\gamma - \theta - \frac{c(i+1)}{\mu} \right]$$

where $\hat{\lambda}$ is the probability that an individual arrives for service during the period for which his right is valid; π_i is the probability that the queue is of size i when he arrives; θ is the toll payable if service is rendered; and n is the queue size at which a customer balks, given the toll θ . Expected revenue per customer is

$$(6) \quad \hat{\lambda} \sum_{i=0}^{n-1} \pi_i \left[\tau + c\gamma - \theta - \frac{c(i+1)}{\mu} \right] + \hat{\lambda}\theta \sum_{i=0}^{n-1} \pi_i \\ = \hat{\lambda} \sum_{i=0}^{n-1} \pi_i \left[\tau + c\gamma - \frac{c(i+1)}{\mu} \right].$$

Due to the constancy of c , revenue per customer is independent of θ except through n . The server's total revenue per time period is revenue per customer times the number of potential customers, N . Note that $\hat{\lambda}N = \lambda$, since the probability of an individual arriving times the number of individuals equals the gross accession rate. The server's objective is to select a critical queue size, by selecting a toll θ , which maximizes

$$(7) \quad \lambda \sum_{i=0}^{n-1} \pi_i \left[\tau + c\gamma - \frac{c(i+1)}{\mu} \right].$$

But (7) is identical to Naor's social welfare function, since $\lambda \sum_{i=0}^{n-1} \pi_i$ is the expected flow of accommodated arrivals;

$$\frac{\lambda c}{\mu} \sum_{i=0}^{n-1} \pi_i (i+1) = c \sum_{i=0}^{n-1} \rho \pi_i (i+1) = c \sum_{i=0}^{n-1} \pi_{i+1} (i+1) \\ = c \sum_{i=1}^n i \pi_i = cE[q|n].$$

That a revenue maximizing entrepreneur will select the socially optimal value for n is another example of the proposition on two-part tariffs stated by Oi [4].

² This two-part tariff is analogous to the scheme proposed in Edelson's deterministic model [1, p. 874, footnote 3].

A single price is charged for the right to service because all customers are identical, and θ in (5) is the toll which sustains a socially optimal balk decision as a private equilibrium.

3. $M/M/1$ MODEL WITH BALKING AND DIFFERENT VALUES FOR c

Different values for c can arise because customers ascribe different costs to waiting or because an individual customer makes repeated arrivals with a diminishing marginal value for service. Heterogeneity in c adds a significant dimension to the balking problem, since to determine θ_s and θ_r one must calculate the expected numbers of waiting customers of each type, not just the expected queue size. To our knowledge, this problem has not previously been considered.

Let $c_1 > c_2$ be the cost of time for two customer types, λ_1 and λ_2 their respective arrival rates, and $\mu_1 = \mu_2 = \mu$ their common service rate. The queue size at which a customer of type i balks, n_i , is defined by the inequalities

$$(8) \quad \mu \left(\frac{\tau - \theta}{c_i} + \gamma \right) \geq n_i > \mu \left(\frac{\tau - \theta}{c_i} + \gamma \right) - 1.$$

Since the critical queue size for the customer with the higher (lower) time cost is less (more) sensitive to changes in the money toll, $n_1 \geq n_2$ as $\theta \geq \tau$. The higher the toll, the greater the expected proportion of type 1 customers in the queue. If $c_1 - c_2$ and $\lambda_1 - \lambda_2$ are large enough, an optimal solution may require $n_2 = 0$, i.e., $\theta \geq \tau + c_2\gamma$, so low-time-value individuals never patronize the facility. Note, however, that if the same toll must be charged to all customers, certain (n_1, n_2) pairs are inconsistent with individual maximization.

$$(9) \quad P = (\lambda_1 - \zeta_1)(\tau + c_1\gamma) + (\lambda_2 - \zeta_2)(\tau + c_2\gamma) - c_1E[q_1; \theta] - c_2E[q_2; \theta]$$

where $\zeta_i = \lambda_i \text{prob} \{q \geq n_i\}$, n_i are determined by (8), and $E[q_i; \theta]$ is the expected number of type i customers in the queue given a toll θ . If no discrimination between customer types is possible, expected revenue per unit time is

$$(10) \quad M = [(\lambda_1 - \zeta_1) + (\lambda_2 - \zeta_2)]\theta$$

where

$$\theta = \tau + c_1 \left(\gamma - \frac{n_1}{\mu} \right) = \tau + c_2 \left(\gamma - \frac{n_2}{\mu} \right).$$

To determine the $E[q_i; \theta]$ we introduce an “indicator function,” a concept familiar to mathematical probabilists but one not widely exploited in applied work. For simplicity, we consider only two customer types and relegate the general case to Appendix C. The indicator function for position k at time t is defined as

$$(11) \quad I_k(t) = \begin{cases} 1 & \text{if position } k \text{ is occupied by a customer of type 1 at time } t, \\ 0 & \text{if position } k \text{ is unoccupied or occupied by a customer of type 2 at time } t. \end{cases}$$

Positions in the system are numbered as follows: 0 is the server, while places in the queue go from 1 (next in line for service) to $n_2 - 1$ (the queue size at which all arrivals balk, given $\theta > \tau$).³

Since type 1 customers balk if $q \geq n_1$, $I_{n_1}(t), I_{n_1+1}(t), \dots$ are automatically zero. Consequently, the total number of type 1 customers in the system at time t , $q_1(t)$, is simply

$$q_1(t) = \sum_{k=0}^{n_1-1} I_k(t).$$

To find $E[q_1(t); \theta]$ we need only evaluate $E[I_k(t)]$ for $k = 0, 1, \dots, n_1 - 1$. The expression $E[q_2(t); \theta]$ is determined as the difference between expected queue size, $E[q; \theta]$, and $E[q_1; \theta]$.

The proof is based on the fact that $q(t)$ and $(q(t), I_0(t), \dots, I_{n_1-1}(t))$ are continuous-time Markov processes. We first write equations for $E[I_k(t + h)]$ conditional upon $I_k(t)$ and $q(t)$, which lead to differential equations in standard fashion. Solving for the steady state solution and taking expectations over $q(t)$ we obtain $E[I_k(t)] = E[I_k]$. For the two category case, the proof in Appendix B demonstrates that

$$(12a) \quad E[q_1; \theta] = \rho_1 \sum_{k=0}^{n_1-1} (k + 1)\rho^k \pi_0$$

$$= \pi_0 \rho_1 (1 - \rho)^{-1} [(1 - \rho)^{-1} (1 - \rho)^{n_1} - n_1 \rho^{n_1}],$$

$$(12b) \quad E[q; \theta] = \pi_0 \{ (1 - \rho)^{-2} \rho F(\rho, n) + \rho^{n_1} (1 - \rho_2)^{-2} \rho_2 G(\rho_2, n_1, n_2) \},$$

$$(12c) \quad E[q_2; \theta] = E[q; \theta] - E[q_1; \theta],$$

where

$$\rho_i = \lambda_i / \mu, \quad i = 1, 2,$$

$$\rho = \rho_1 + \rho_2 = \frac{\lambda_1 + \lambda_2}{\mu},$$

$$\pi_0 = \left[\sum_{k=0}^{n_1-1} \rho^k + \rho^{n_1} \sum_{k=n_1+1}^{n_2} \rho_2^{k-n_1} \right]^{-1}$$

(probability that there are no customers in the system),

$$F(\rho, n_1) = 1 - \rho^{n_1} - (1 - \rho)n_1 \rho^{n_1},$$

$$G(\rho_2, n_1, n_2) = (1 - \rho_2^{n_2-n_1})(1 + n_1 - n_1 \rho_2) - (n_2 - n_1)(1 - \rho_2)\rho_2^{n_2-n_1}.$$

With only two customer types P , the social welfare function, is unimodal in θ but M , the revenue function, is bimodal. As in a deterministic model θ_r can be greater than, equal to, or less than θ_s if $c_1 \neq c_2$. The paradoxical case, $\theta_r < \theta_s$, occurs if c_1/c_2 is large relative to λ_2/λ_1 , but not so large that the optimal toll

³ When $\theta < \tau$, the maximal queue length is $n_1 - 1$, so the analysis is identical with just a change in notation.

completely excludes low-time-value customers. For example, let $\mu = 1$, $\tau = 100$, and $\gamma = 3$. When $\lambda_1 = 0.35$, $\lambda_2 = 1.00$, $c_1 = 100$, and $c_2 = 10$, it can be shown that $\theta_r = 100 < \theta_s = 120$. With these parameter values price is too low even if direct price discrimination is feasible. The Pareto optimal pair of tolls is $(\theta_s^1 = 200, \theta_s^2 = 120)$, where $(\theta_r^1 = 200, \theta_r^2 = 110)$. If λ_1 decreases to 0.25 and c_2 increases to 15, $\theta_r = \theta_s = 115$. The usual monopoly result, $\theta_r > \theta_s$, occurs with the above parameter values if c_1 falls from 100 to 50. Curiously, $\theta_r = \theta_s$ for a wide range of values for c_1 if $c_2 = 10$. In all cases where we found $\theta_r \neq \theta_s$, the percentage loss in expected welfare was small.

Although a two-part tariff is Pareto optimal when customers have the same value for time, this is not necessarily the case when $c_1 \neq c_2$. In fact, a revenue maximizing two-part tariff will produce a lower level of expected social welfare than a single toll if $\theta_r < \theta_s$ and $\theta \leq \theta_r$. The latter inequality is satisfied if the minimum offer for a service option comes from low-time-value customers,⁴ a condition which is met if the arrival rates are high enough to ensure that θ_r is not much smaller than τ . Consider the parameter set $\lambda_1 = 0.25$, $\lambda_2 = 1.00$, $c_1 = 100$, and $c_2 = 8$, for which $\theta_r = 100$ and $\theta_s = 108$. With a two-part tariff the price for a service option is 7.32 (a high-time-value customer offers 46.04), and the charge for service is 92. Expected social welfare at this equilibrium is only 98.91, as compared with 101.26 at $\theta_r = 100$. It is also possible to construct examples where $\theta_s = \theta_r$, but the charge for service under a two-part tariff is less than θ_s .

Another interesting question is whether a monopolist will install the socially optimal number of servers. Since a different price can be charged at each service facility, expanding the number of facilities is a way to segment the market. It is often claimed that goods provided through a political process are overly standardized, since in the absence of logrolling a majority coalition will impose its preference on the rest of society. The opportunity for practicing price discrimination suggests that an opposite bias may exist in private markets.

There is a situation where a monopolist will necessarily install the socially optimal number of servers and charge the socially optimal toll at each facility. Suppose that Pareto optimality requires at least one server for each customer type, and that under a two-part tariff no customer finds it worthwhile to affiliate with more than one server. It is clear that a profit maximizer will segment the market and,

⁴ Expected revenue under a two-part tariff is, if both customer types purchase the service option,

$$(\lambda_1 + \lambda_2) \left\{ \min_{i=1,2} \sum_{j=0}^{n_i-1} \pi_j \left[\tau + c_i \gamma - \theta - \frac{c_i(j+1)}{\mu} \right] \right\} + \theta \left[\lambda_1 \sum_{j=0}^{n_1-1} \pi_j + \lambda_2 \sum_{j=0}^{n_2-1} \pi_j \right],$$

where the n_i are determined by θ in the usual fashion. At small values for θ high-time-value customers determine the price of a service option, since it is they who have the smaller consumer surpluses, but before $\theta = \tau$ low-time-value customers become decisive. Consumer surplus for a low-time-value customer is a decreasing function of θ , whereas the opposite is true for a high-time-value customer if $\lambda_1 < \lambda_2$. Therefore, a sufficient condition for $\theta \leq \theta_r$ is that at θ_r the service option price be determined by the low-time-value customers. If this is so, increasing θ above θ_r decreases both the toll component of expected revenue and the proceeds from sale of service options. This leaves open the possibility that $\theta > \theta_r$ if θ_r is small enough so that the service option price is set by high-time-value customers. We were not able to construct such a case, however. The outcome $\theta > \theta_r$ seems unlikely because in a deterministic model, the price per unit of output under a two-part tariff must be less than the single monopoly price [4, Equation 10].

invoking the result from Section 2, the two-part tariff at each facility will be Pareto optimal.

If the entrepreneur is restricted to charging simple tolls, however, it is possible that he will invest in an excessive number of servers. Suppose that with two customer types and two servers, high-time-value customers patronize only the high toll facility and vice versa.⁵ Consider the parameter set $\lambda_1 = 0.35$, $\lambda_2 = 1.00$, $c_1 = 100$, and $c_2 = 10$. With a single server $\theta_r = 100 < \theta_s = 120$, and at the revenue maximizing toll expected revenue is 84.92 and expected social welfare 107.65. If there are two specialized servers, expected revenue is maximized at 152.78 if $\theta_1 = 200$ and $\theta_2 = 100$. At these tolls expected social welfare is 160.28. The increase in expected revenue, $152.78 - 84.92 = 67.86$, exceeds the increase in expected social welfare at the revenue maximizing tolls, $160.28 - 107.65 = 52.63$.

Although it would be profitable for a monopolist to install a second server if its cost per unit time were, say, 60, from the point of view of Pareto optimality that investment should not be undertaken. Investment in a second server may be privately profitable but socially undesirable even if, with a single server, $\theta_r = \theta_s$. The paradox arises, of course, because assigning property rights to a monopolist places the problem in a second-best context. If Pareto optimal tolls were charged on the two specialized facilities, as would be the case with a two-part tariff, a privately profitable investment would also increase expected social welfare.

We reach somewhat different judgements about the provision of capacity if, for legal or administrative reasons, the same toll must be charged at each facility. Suppose that when there are z servers the arrival rate of customer type i at each server is λ_i/z .⁶ Numerical examples show that a profit maximizing monopolist is likely to install fewer than the optimal number of servers, since, except at isolated points, the change in expected revenue is less than the change in expected welfare.

On the other hand, too many servers will exist if there is free entry subject to the requirement that every producer charge the prevailing toll. Excess capacity arises because each potential entrepreneur compares his share of industry expected revenue when there are $(n + 1)$ servers with long run marginal cost, but the change in expected welfare is less than expected revenue per facility at the optimum number of servers. Note that a similar divergence between average private cost and marginal social cost is what justifies imposing a congestion toll on "peak-load" arrivals.

⁵ We were unable to solve for the steady state solution of a model where an arriving customer can scan two queues and join either one of them or balk to a third server represented by (τ, γ) . An example of this situation is a restaurant offering sit-down or take-out service. Although a simple phase diagram can be drawn, expressions for the transition probabilities are quite complex around the balk point.

⁶ In this model expected welfare (revenue) is obtained by multiplying expected welfare (revenue) per server, where the arrival rates are λ_i/z , by the number of servers. One complication is that if c_1 is much larger than c_2 , neither expected welfare nor expected revenue is a unimodal function of n ; in some instances the change in expected revenue exceeds the change in expected welfare as the number of servers is augmented by one. Moreover, in such cases the revenue maximizing toll sustains the balking rule ($n_1 = 1, n_2 = 0$) when there are few servers, but θ_r decreases sharply at some critical value for z . The Pareto optimal toll is a more smoothly decreasing function of z .

4. CONCLUSIONS

Incorporating heterogeneous values for waiting time into Naor's model is, in one sense, a disappointing exercise: one loses the implications that θ_s is no larger than θ_r and that a two-part tariff is Pareto optimal without getting other concrete results in return. It is interesting to note that $\theta_r = \theta_s$ for a wide range of parameter values, and that when $\theta_r \neq \theta_s$ the percentage loss in expected welfare is quite small. In addition, the model provides insights into the relationship between price discrimination and excess capacity, and the benefit of charging two-part tariffs when facilities service a relatively homogeneous clientele.

A useful technical contribution of the paper is its application of the indicator function technique to queuing theory. There appear to be many practical problems where a dichotomous classification of customers is an acceptable approximation, e.g., emergency vs. non-emergency medical cases, and in this situation the indicator function can easily determine an optimal configuration of tolls and servers.

University of Pennsylvania

Manuscript received February, 1973; revision received November, 1973.

APPENDIX A

A QUEUING MODEL WITHOUT BALKING WHERE THE ARRIVAL RATE DEPENDS ON TOLLS

Order the population of potential customers in decreasing order with respect to their values for time. For any toll $\theta > \tau$, that segment of the population having value of time $c_L(\theta)$ will be indifferent between the two service facilities, where $c_L(\theta)$ is defined by

$$(A1) \quad \tau + \gamma c_L(\theta) = \theta + c_L(\theta)E[t; \theta].$$

All customers with $c > c_L(\theta)$ prefer the θ -facility and vice-versa. For $\theta < \tau$ there exists a $c_U(\theta)$ such that

$$(A2) \quad \tau + \gamma c_U(\theta) = \theta + c_U(\theta)E[t; \theta],$$

and all customers with $c < c_U(\theta)$ prefer the θ -facility. If $\theta > \tau$ ($\theta < \tau$) raising θ causes customers with the lowest (highest) values for time from among those patronizing the θ -facility to be diverted to the τ -facility.

Let $\bar{c}_L(\theta) > c_L(\theta)$ be the average value for customers patronizing the θ -facility when $\theta > \tau$, and $\bar{c}_U(\theta) < c_U(\theta)$ when $\theta < \tau$. The expression for expected social welfare is the same as (3) except that $\bar{c}_L(\theta)$ replaces c when $\theta > \tau$ and $\bar{c}_U(\theta)$ replaces c when $\theta < \tau$. Average revenue per customer, θ , is defined by (A1) and (A2), respectively.

It can be seen why θ_r may be less than θ_s if $\theta_s > \tau$. Increasing expected customer flow by lowering θ increases expected waiting time, and the social welfare function evaluates this added delay in terms of $\bar{c}_L(\theta)$. The private entrepreneur considers only the amount by which he must lower θ , and in (A1) $E[t; \theta]$ is multiplied by $c_L(\theta) < \bar{c}_L(\theta)$. The value of θ_r can be less than θ_s if the tendency to undervalue additional delays is large enough relative to the revenue maximizer's incentive to raise θ in order to expropriate inframarginal benefits. The inequality $\theta_r < \theta_s < \tau$ cannot occur because $c_U(\theta) > \bar{c}_U(\theta)$, i.e., the private entrepreneur overvalues the social cost of additional delays.

APPENDIX B

EXPECTED QUEUE COMPOSITION FOR THE TWO-CATEGORY CASE

Since the system reduces to Naor's queueing process if $n_1 = 0$, let $n_2 > n_1 > 0$. For h small and neglecting terms of order h^2 and higher, the probability of a type 1 arrival is $\lambda_1 h$, while the probability of service completion, given $q(t) \geq 1$, is μh . Therefore,

(B1) given $q(t) = 0$,

$$E[I_0(t + h)] = \text{prob} \{I_0(t + h) = 1\} \cong \lambda_1 h;$$

(B2) given $q(t) = 1$ and $I_0(t)$, all other $I_k(t)$ being of necessity zero,

$$\begin{aligned} E[I_0(t + h)] &\cong (1 - \mu h)I_0(t) \\ &= (1 - \mu h)I_0(t) + \mu h I_1(t), \quad \text{since } I_1(t) = 0, \quad \text{and} \\ E[I_1(t + h)] &\cong \lambda_1 h. \end{aligned}$$

Thus, if $I_0(t) = 1$, $E[I_0(t + h)]$ is the probability of no service completion; if $I_0(t) = 0$, which means a type 2 customer is in service, the probability that $I_0(t + h) = 1$ is of order h^2 .

(B3) Given $q(t) = K, 2 \leq K < n_1 - 2$, and $I_0(t), I_1(t), \dots, I_{K-1}(t)$,

$$\begin{aligned} E[I_m(t + h)] &\cong (1 - \mu h)I_m(t) + \mu h I_{m+1}(t) && (m = 0, 1, \dots, K - 1), \\ E[I_K(t)] &\cong \lambda_1 h; \end{aligned}$$

(B4) given $q(t) = K$ and $n_1 - 1 \leq K \leq n_2 - 1$, since balking occurs,

$$\begin{aligned} E[I_m(t + h)] &\cong (1 - \mu h)I_m(t) + \mu h I_{m+1}(t) && (m = 0, 1, \dots, n_1 - 1), \\ E[I_m(t + h)] &= 0 && (m = n_1, \dots, n_2 - 1). \end{aligned}$$

Next, take expectations over the $I_m(t)$, still conditional on $q(t)$. The left-hand sides above become $E[I_m(t + h)]$ given $q(t)$ alone, while the right-hand sides have $I_m(t)$ and $I_{m+1}(t)$ replaced by their expectations given $q(t)$. Suppose that the process is in statistical equilibrium, so that the $\pi_K = \text{prob} \{q(t) = K\}$ are constant over time. Multiplying by these π_K and adding, i.e., taking expectations over $q(t)$, we obtain

$$\begin{aligned} E[I_0(t + h)] &\cong \lambda_1 h \pi_0 + \sum_{K=1}^{n_1-1} (1 - \mu h) E[I_0(t) | q(t) = K] \pi_K + \sum_{K=1}^{n_1-1} \mu h E[I_1(t) | q(t) = K] \pi_K \\ &= \lambda_1 h \pi_0 + (1 - \mu h) E[I_0(t)] + \mu h E[I_1(t)]; \end{aligned}$$

similarly, for $1 \leq m \leq n_1 - 2$. Note that for $m = n_1 - 1$ we have

$$E[I_m(t + h)] = \lambda_1 h \pi_{n_1-1} + (1 - \mu h) E[I_{n_1-1}(t)].$$

Now form the usual differential equations by taking the limit of

$$\frac{1}{h} \{E[I_m(t + h)] - E[I_m(t)]\}$$

as $h \rightarrow 0$. These derivatives are all zero if the system is in statistical equilibrium, and we solve to get

$$\begin{aligned} E[I_{n_1-1}(t)] &= \frac{\lambda_1}{\mu} \pi_{n_1-1} = \rho_1 \pi_{n_1-1} \quad \text{and} \\ E[I_m(t)] &= \frac{\lambda_1}{\mu} \pi_m + E[I_{m+1}(t)] && (0 \leq m \leq n_1 - 2) \\ &= \rho_1 (\pi_m + \dots + \pi_{n_1-1}), \quad \text{by induction.} \end{aligned}$$

Therefore,

$$E[q_1; \theta] = \sum_{m=0}^{n_1-1} E[I_m(t)] = \rho_1 \sum_{K=0}^{n_1-1} (K + 1) \pi_K.$$

Evaluation of the π_K is straightforward, since $q(t)$ is a simple Markov process. By the usual process of solving differential equations in the probabilities we find

$$\begin{aligned} \pi_K &= \rho^K \pi_0 & (K = 1, 2, \dots, n_1), \\ \pi_K &= \rho^{n_1} \rho_2^{K-n_1} \pi_0 & (K = n_1 + 1, \dots, n_2). \end{aligned}$$

Therefore,

$$\pi_0 = \left\{ \sum_{K=0}^{n_1-1} \rho^K + \rho^{n_1} \sum_{K=n_1+1}^{n_2} \rho_2^{K-n_1} \right\}^{-1}.$$

Of course, the formula for the probability of an empty queue, π_0 , assumes that ρ_2 and ρ are not unity. The obvious limits exist when $\rho_2 \rightarrow 1$ or $\rho \rightarrow 1$.

Substituting the values for π_K yield the expressions (12a), (12b), and (12c) in the text.

APPENDIX C

EXPECTED QUEUE COMPOSITION FOR THE MULTI-CATEGORY CASE

Order the C customer types such that $n_j > n_{j-1}$. The indicator function for position K at time t is

$$I_{j,K}(t) = \begin{cases} 1 & \text{if a customer of type } j \text{ is in position } K \text{ at time } t, \\ 0 & \text{otherwise.} \end{cases}$$

Given $q(t) \leq K - 1$ ($1 \leq K \leq m_j - 1$),

$$E[I_{j,K}(t + h)] = o(h).$$

Given $q(t) = K$ and $\{I_{j,K}(t)\}$,

$$E[I_{j,K}(t + h)] = \lambda_j h + o(h).$$

Given $q(t) \geq K + 1$ and $\{I_{j,K}(t)\}$,

$$E[I_{j,K}(t + h)] = (1 - \mu h)I_{j,K}(t) + \mu h I_{j,K+1}(t) + o(h).$$

Given $q(t) = 0$, and $K = 0$,

$$E[I_{j,0}(t + h)] = \lambda_j h + o(h).$$

Given $q(t) \geq 1$ and $K = 0$,

$$E[I_{j,0}(t + h)] = (1 - \mu h)I_{j,0}(t) + \mu h I_{j,1}(t) + o(h).$$

First take conditional expectations, then expectations over $q(t)$:

$$E[I_{j,K}(t + h)] = \pi_K \lambda_j h + (1 - \mu h)E[I_{j,K}(t)] + \mu h E[I_{j,K+1}(t)] + o(h), \quad K \leq m_j - 1.$$

The differential equations in statistical equilibrium yield

$$E[I_{j,K}(t)] = \pi_K \frac{\lambda_j}{\mu} + E[I_{j,K+1}(t)]$$

as in the two category case. Since $E[I_{j,m_j}(t)] = 0$,

$$E[I_{j,K}(t)] = \frac{\lambda_j}{\mu} \sum_{K'=K}^{m_j-1} \pi_{K'}, \quad 0 \leq K \leq m_j - 1.$$

The π_K are easily obtained because $q(t)$ is a Markov process,

$$\text{prob } \{q(t + h) = 0\} = \left(1 - \sum_{j=1}^C \lambda_j h \right) \text{prob } \{q(t) = 0\} + \mu h \text{prob } \{q(t) = 1\} + o(h).$$

Therefore,

$$\pi_1 = \frac{1}{\mu} \eta_1 \pi_0, \quad \text{where } \eta_1 = \sum_{j=1}^C \lambda_j,$$

and for $K \leq m_1 - 1$

$$\begin{aligned} \text{prob } \{q(t+h) = K - 1\} &= \eta_1 h \text{ prob } \{q(t) = K - 2\} + \mu h \text{ prob } \{q(t) = K\} \\ &+ (1 - \eta_1 h - \mu h) \text{ prob } \{q(t) = K - 1\}. \end{aligned}$$

By the usual manipulations,

$$\pi_K = \pi_{K-1} + \frac{1}{\mu} \eta_1 \pi_{K-1} - \frac{1}{\mu} \eta_1 \pi_{K-2}.$$

Assume inductively that $\pi_{K-1} = (1/\mu)\eta_1\pi_{K-2}$. Then $\pi_K = (1/\mu)\eta_1\pi_{K-1}$, confirming the induction. For larger K , the only difference is the arrival rate. The same analysis yields

$$\pi_K = \frac{1}{\mu} \eta_j \pi_{K-1}, \quad m_{j-1} \leq K \leq m_j - 1,$$

where $\eta_j = \sum_{j'=j}^C \lambda_{j'}$. A more explicit solution is as follows:

$$\pi_K = \left(\frac{\eta_j}{\mu}\right)^{K-m_{j-1}} \left(\frac{\eta_{j-1}}{\mu}\right)^{m_{j-1}-m_{j-2}} \dots \left(\frac{\eta_1}{\mu}\right)^{m_1-0} \pi_0, \quad m_{j-1} \leq K \leq m_j - 1.$$

REFERENCES

[1] EDELSON, NOEL M.: "Congestion Tolls Under Monopoly," *American Economic Review*, 59 (1971), 873-882.
 [2] LITTLE, JOHN D. C.: "A Proof for the Queuing Formula: $L = \lambda W$," *Operations Research*, 9 (1961), 383-387.
 [3] NAOR, P.: "The Regulation of Queue Size By Levying Tolls," *Econometrica*, 37 (1969), 15-24.
 [4] OI, WALTER Y.: "A Disneyland Dilemma: Two-Part Tariffs for a Mickey Mouse Monopoly," *Quarterly Journal of Economics*, 85 (1971), 77-96.
 [5] YECHIALI, URI: "On Optimal Balking Rules and Toll Charges in the $GI/M/1$ Queueing Process," *Operations Research*, 19 (1971), 349-370.