Econometrica, Vol. 37, No. 1 (January, 1969)

THE REGULATION OF QUEUE SIZE BY LEVYING TOLLS[1]

BY P. NAOR

A queueing model—together with a cost structure—is presented, which envisages the imposition of tolls on newly arriving customers. It is shown that frequently this is a strategy which might lead to the attainment of social optimality.

1. INTRODUCTION

SOME DISCUSSION has arisen recently as to whether the imposition of an "entrance fee" on arriving customers who wish to be serviced by a station—and hence join a waiting line—is a rational measure. Not much of this discussion has appeared in print; indeed this author is aware of only three short communications, representing an exchange of arguments between Leeman [**1**, **2**] and Saaty [**3**]. The ideas advanced there were of qualitative character and no attempt was made to quantify the arguments.

The problem under consideration is obviously analogous to one that arises in connection with the control of vehicular traffic congestion on a road network. It has been argued[2] by traffic economists that the individual car driver—on making an optimal routing choice for himself—does not optimize the system at large. The purpose of this communication is to demonstrate that, indeed, analogous conclusions can be drawn for queueing models if two basic conditions are satisfied:

CONDITION I: A public good is identifiable for which an objective function (typically a profit function) can be set up and maximized. This state of affairs—the existence of a public good—may manifest itself in several distinct ways, two of which are as follows: (a) The population of arriving customers and the service station(s) are under the control of a single decision maker who represents the public good. (b) The arriving customers represent decision makers and "everybody is in business for himself." Utilities to these distinct decision makers are comparable and additive, however, and gains may be redistributed, e.g., through the agency of a mutual risk insurance company. Hence the expected overall profit (in unit time) accruing to arriving customers is a proper objective function representing public good.

CONDITION II: Customers are liable to be diverted from the service station.

[1] This research was supported by the Office of Naval Research under Contract No. Nonr-855(09).

[2] This author had the privilege of attending a Colloquium on "Decision Making in Traffic Planning" organized in summer 1965 by Professor Arne Jensen of the Technical University, Copenhagen. Professor Martin Beckmann of Brown University and Bonn University, in his lecture at that Colloquium, presented convincing arguments in favor of the thesis that the routing decision of the individual driver optimizing his own interest will not typically optimize an overall objective function. Hence imposing appropriate tolls may bring about optimal redistribution of vehicles moving within the road network.

In other words, some of them will be directed not to queue up (and not to invest time in that process) and not to reap the benefits available through the service at the station. This condition is in striking contrast to the usual assumptions made in queueing situations and therefore it is thought useful to elaborate a little on this point. Typically, it is assumed in most models that the mission of the station is to render service to *all* arriving customers (if only this does not violate the steady state condition) and an amelioration of congestion (that is: cutting of losses) may be brought about by sequencing the customers in some prescribed order. This is the rationale of most priority queueing models. In the present model the sole means of control at the disposition of the decision maker is the possible nonadmission of the newly arrived customer to the waiting line. The purpose of a toll (or an equivalent administrative measure) is precisely to prevent customers from joining the queue in case of heavy congestion, and without the present condition—"customers are liable to be diverted from the service station"—there can be no rationale for the levying of tolls.

Having posed the above fundamental conditions necessary to create a framework for a queueing system with tolls we can now proceed to a detailed description of a model and a cost structure:

(i) A stationary Poisson stream of customers—with parameter λ—arrives at a single service station.

(ii) The station renders service in such a way that the service times are independently, identically, and exponentially distributed with intensity parameter μ.

(iii) On successful completion of service, the customer is endowed with a reward R (expressible in monetary units). All customer rewards are equal.

(iv) The cost to a customer for staying in a queue (i.e. for queueing) is C monetary units in unit time. All customer costs are equal.

(v) The newly arrived customer is required to choose one of two alternatives: either (a) he joins the queue, incurs the losses associated with spending some of his time in it, and finally obtains the reward; or (b) he refuses to join the queue—an action which does not bring about any gain or loss. The choice of one of these two alternatives will be made by the customer on comparing the net gains associated with each of them. To avoid ambiguity it will be stipulated that, in the case of a tie, the customer will join the queue.

It is immediately clear that some of these assumptions represent gross simplifications of "real life" and cannot ordinarily be asserted to faithfully represent reality. Thus, for instance, there is no reason to assume that in "real life" service times are exponentially distributed, that rewards to all customers are equal (rather than statistically distributed), that queueing expenditure per unit time is identical for all customers, etc. These specific assumptions were made here since they facilitate mathematical manipulation without needlessly obscuring structure. A pertinent feature of our model—rather easily demonstrated on making the present

specific assumptions—is the following: *Exercise of narrow self interest by all customers does not optimize public good.* This feature of the model is preserved under generalizations, e.g., if no specific assumptions as to the nature of the service process are made.

Finally, in this section, two existing characteristics of our model are stated. First, it is not necessary to make the assumption—usual in most queueing models with a single service station—that, for steady state conditions to exist, the service intensity μ must exceed the arrival intensity λ. Arriving customers are liable to be diverted from the service station in the present model so that what is required for steady state conditions to prevail is only that the average number (per unit time) of nondiverted customers must fall short of service capacity μ. This will always be the case under the model assumptions enumerated before.

Secondly, for our model to make sense it is required that under favorable circumstances a customer will desire to queue up for (completion of) service. Hence if a newly arrived customer encounters a completely empty service station he should make the "pro-queueing" choice. His expected loss is given by $C\mu^{-1}$ whereas his reward to be collected at the end of service equals R. If the dimensionless quantity $(R\mu/C)$ is denoted by v_s it is clear that in meaningful models the following inequality must hold

$$(1) \qquad v_s = \frac{R\mu}{C} \geqslant 1 .$$

If inequality (1) does not hold the optimal policy is to disband the service station and divert the customer stream altogether.

2. SOME PROPERTIES OF THE MODEL

Under the model assumptions given in the previous section, it is clear that all reasonable strategies will be of the following nature. A newly arrived customer will observe the queue size, i say, at that instant. This quantity is a random variable whose distribution is partly determined by the strategy pursued. Now if the observed value of this random variable falls short of a constant n (the selected strategy), the newly arrived customer will join the queue; if the observed value i is equal to n the new customer is diverted and does not join the queue. The observed value i can never exceed n in this model.

Clearly we are confronted with a system that is identical with a queueing model in which finite waiting space only is available to queueing customers. If we define

$$(2) \qquad \frac{\lambda}{\mu} = \rho ,$$

we obtain the following steady state equations:

$$(3) \qquad p_i\rho = p_{i+1} \qquad (0 \leqslant i < n),$$

the solution[3] of which is

$$p_i = \frac{\rho^i}{1+\rho+\ldots+\rho^n} = \frac{\rho^i(1-\rho)}{1-\rho^{n+1}} \qquad (0 \leqslant i \leqslant n). \tag{4}$$

The generating function is derived as

$$g(z) = \sum_{i=0}^{n} p_i z^i = \frac{1-\rho}{1-\rho^{n+1}} \cdot \frac{1-(\rho z)^{n+1}}{1-\rho z}. \tag{5}$$

The expected value, q, of the random variable equals

$$q = E\{i\} = \frac{\rho[1-(n+1)\rho^n+n\rho^{n+1}]}{(1-\rho)(1-\rho^{n+1})} = \frac{\rho}{1-\rho} - \frac{(n+1)\rho^{n+1}}{1-\rho^{n+1}}. \tag{6}$$

The expected number of customers, ζ say, diverted from the service station in unit time is given by

$$\zeta = \lambda p_n = \frac{\lambda\rho^n(1-\rho)}{1-\rho^{n+1}}. \tag{7}$$

We mention, in passing, that the busy fraction b, i.e., the degree of utilization of the service station, is, of course, not equal to ρ (as in the "usual" models) but rather

$$b = \sum_{i=1}^{n} p_i = 1 - p_0 = \frac{\rho(1-\rho^n)}{1-\rho^{n+1}}. \tag{8}$$

The expected number of customers joining the queue in unit time equals

$$\lambda - \zeta = \lambda(1-p_n) = \lambda\left[1 - \frac{\rho^n(1-\rho)}{1-\rho^{n+1}}\right] = \lambda\frac{1-\rho^n}{1-\rho^{n+1}}. \tag{9}$$

The expected number of customers leaving the service station in unit time equals

$$\mu b = \mu(1-p_0) = \mu\left[1 - \frac{1-\rho}{1-\rho^{n+1}}\right]. \tag{10}$$

These two quantities must be identical under steady state conditions and, indeed, it is easy to verify that

$$\rho = \frac{1-p_0}{1-p_n}. \tag{11}$$

3. SELF-OPTIMIZATION

Let us now assume that a strategy (which will be designated as n_s) is selected in the following manner (envisaged already in a general way in the Introduction):

[3] Throughout this study functions of ρ will make their appearance which generate indeterminate forms—0/0 or $\infty - \infty$—on insertion of the value $\rho = 1$; in all cases a unique (nonzero and finite) limit exists. Hence we shall omit notice that the formulas are valid only if $\rho \neq 1$.

The newly arrived customer weighs the two alternatives—to join or not to join the queue—by the net gains associated with them. The net gain, in the first case, is equal to

$$G_i = R-(i+1)C\frac{1}{\mu}. \tag{12}$$

In the alternative case the net gain is zero. Hence self interest is served if a strategy is established in the following fashion. An integer, n_s, is found which satisfies simultaneously two inequalities

$$R-n_sC\frac{1}{\mu} \geqslant 0 \tag{13}$$

and

$$R-(n_s+1)C\frac{1}{\mu} < 0\,. \tag{14}$$

Inequality (13) pertains to the case where the number of queueing customers (including the one in service) encountered by the newly arrived customer falls short of the critical number by one. The customer, of course, is supposed to join and indeed the inequality is in his favor. Inequality (14) relates to the unfavorable event: the critical number, n_s, of customers is already in the queue. We can incorporate the two inequalities in one expression

$$n_s \leqslant \frac{R\mu}{C} = v_s < n_s+1\,. \tag{15}$$

Alternatively we may express the same idea in different notation

$$n_s = [v_s] \tag{16}$$

where [ ] is the well known bracket function; that is, n_s is the largest integer not exceeding v_s.

We note that the critical number, n_s, derived by "actualizing self interest" depends on μ, R, and C, but not on the arrival intensity λ. This fact alone suffices—before pursuing further detailed investigation—to throw serious doubt on the social optimality of the strategy n_s.

4. OVERALL OPTIMIZATION

If the viewpoint is taken that the expected sum of the net gains accruing to customers in unit time is the public good which should be optimized, we must proceed in a different mode from that outlined in the previous section. We note that expected total net gain, P, under some strategy n is given by

$$\begin{aligned} P &= (\lambda-\zeta)R-CE\{i\} = \lambda R(1-p_n)-Cq \\ &= \lambda R\frac{1-\rho^n}{1-\rho^{n+1}} - C\left[\frac{\rho}{1-\rho}-\frac{(n+1)\rho^{n+1}}{1-\rho^{n+1}}\right]. \end{aligned} \tag{17}$$

By some elementary (but lengthy) considerations it can be shown that the function P in its dependence on n is "discretely unimodal" or, in other words, a local maximum is a global maximum. Hence we seek that strategy, n_0 say, which is associated with two inequalities

$$\lambda R\left[\frac{\rho^{n_0}(1-\rho)}{1-\rho^{n_0+1}}-\frac{\rho^{n_0+1}(1-\rho)}{1-\rho^{n_0+2}}\right]-C\left[\frac{(n_0+1)\rho^{n_0+1}}{1-\rho^{n_0+1}}-\frac{(n_0+2)\rho^{n_0+2}}{1-\rho^{n_0+2}}\right]<0 \tag{18}$$

and

$$\lambda R\left[\frac{\rho^{n_0-1}(1-\rho)}{1-\rho^{n_0}}-\frac{\rho^{n_0}(1-\rho)}{1-\rho^{n_0+1}}\right]-C\left[\frac{n_0\rho^{n_0}}{1-\rho^{n_0}}-\frac{(n_0+1)\rho^{n_0+1}}{1-\rho^{n_0+1}}\right]\geqslant 0\,. \tag{19}$$

Further manipulations transform (18) and (19) into equivalent inequalities

$$\begin{aligned} R(1-\rho)^2 &< \frac{C}{\mu}\left[1-2\rho+n_0(1-\rho)+\rho^{n_0+2}\right] \\ &= \frac{C}{\mu}\left[(n_0+1)(1-\rho)-\rho(1-\rho^{n_0+1})\right] \end{aligned} \tag{20}$$

and

$$R(1-\rho)^2 \geqslant \frac{C}{\mu}\left[n_0(1-\rho)-\rho(1-\rho^{n_0})\right]\,. \tag{21}$$

These two inequalities in turn can be cast into the following form

$$\frac{n_0(1-\rho)-\rho(1-\rho^{n_0})}{(1-\rho)^2}\leqslant\frac{R\mu}{C}<\frac{(n_0+1)(1-\rho)-\rho(1-\rho^{n_0+1})}{(1-\rho)^2}\,. \tag{22}$$

To deal with (22) it will be convenient to investigate a function,

$$v_s=\left[v_0(1-\rho)-\rho(1-\rho^{v_0})\right](1-\rho)^{-2} \tag{23}$$

of two independent variables $\rho(>0)$ and $v_0(\geqslant 1)$. We note, in passing, that no true singularity exists for this function if $\rho=1$; rather the function is well behaved and a nonzero and finite function value exists at that value of ρ, to wit: $v_s=(v_0(v_0+1)/2)$. Next we study a setting in which the value of ρ is arbitrary (positive) but fixed; we note that v_s is a boundlessly increasing function of v_0. Hence the integers between which v_0 lies (viewed now as a function of v_s and ρ) will obey the inequalities associated with (22). As a result we arrive at

$$n_0=[v_0]\,, \tag{24}$$

an expression which is completely analogous to (18). Further manipulations lead to the inequality

$$v_0\leqslant v_s \tag{25}$$

where the equality sign holds only[4] if v_s equals unity.

[4] The equality sign would hold also in the physically meaningless, and therefore excluded, case $\rho=0$ (arbitrary v_s).

To exhibit some properties of the functional relationship discussed in the present section the form $v_0 = v_0(v_s, \rho)$, i.e., v_0 as a function of the arguments v_s and ρ, was chosen. A set of numerical values of v_0 was computed for selected values of the arguments v_s and ρ. The results are presented in Table I. We note, in passing, that an interesting specific numerical relationship exists: whenever $v_s - \rho = 2$ this results

TABLE I

THE FUNCTION v_0 DEPENDENT ON THE ARGUMENTS v_s AND ρ

v_s	$\rho=0.100$	$\rho=0.200$	$\rho=0.500$	$\rho=1.000$	$\rho=2.000$	$\rho=3.000$	$\rho=4.000$	$\rho=5.000$
1.000	1.000	1.000	1.000	1.000	1.000	1.000		
1.500	1.457	1.425	1.360	1.303	1.247	1.218		
2.000	1.910	1.837	1.690	1.561	1.445	1.387		
2.500	2.361	2.243	2.000	1.791	1.612	1.527		
3.000	2.811	2.647	2.297	2.000	1.756	1.645		
3.500	3.261	3.048	2.583	2.192	1.885	1.750		
4.000	3.711	3.449	2.863	2.372	2.000	1.841	1.749	
4.500	4.161	3.849	3.136	2.541	2.105	1.925	1.821	
5.000	4.611	4.250	3.405	2.702	2.202	2.000	1.886	1.811
5.500	5.061	4.650	3.671	2.854	2.292	2.070	1.945	1.864
6.000	5.511	5.050	3.935	3.000	2.375	2.134	2.000	1.913
6.500	5.961	5.450	4.195	3.140	2.453	2.194	2.051	1.958
7.000	6.411	5.850	4.454	3.275	2.527	2.249	2.098	2.000
7.500	6.861	6.250	4.712	3.405	2.597	2.301	2.142	2.039
8.000	7.311	6.650	4.968	3.531	2.663	2.351	2.184	2.077
8.500	7.761	7.050	5.223	3.653	2.725	2.398	2.223	2.111
9.000	8.211	7.450	5.478	3.772	2.785	2.442	2.260	2.144
9.500	8.661	7.850	5.731	3.887	2.842	2.484	2.295	2.175
10.000	9.111	8.250	5.984	4.000	2.897	2.524	2.328	2.205

in $v_0 = 2$ and the converse of this statement holds as well. It is easy to prove both implications. Indeed one may even state (and prove) the following generalizations: (a) whenever the inequality $v_s - \rho > 2$ holds, the appropriate value of v_0 can be bracketed by $v_s - \rho > v_0 > 2$; and (b) whenever the inequality $v_s - \rho < 2$ holds, it may be established that $v_s - \rho < v_0 < 2$.

5. BENEFICIAL TOLL IMPOSITION

Inequality (25) (which typically would be strict) points to the fact that consideration of narrow self interest does not ordinarily lead to overall optimality. We note, of course, that even a strict inequality need not demonstrate a socially nonoptimal situation if self interest is actualized since both v_s and v_0 may possibly be found between the same integers such that $[v_s]$ and $[v_0]$ are identical. Frequently it should be expected however, that a situation is realized in which—for the sake of

narrow self interest—the facilities of the system are overly congested. To arrive at an ameliorated state of affairs it is necessary to reduce the strategy n from n_s to n_0. This can be done in two distinct ways: either through an administrative rule to the effect that the maximally permissible queue size should be smaller[5] than a prima facie admissible number n_s; or, alternatively, a toll θ is imposed on customers joining the queue and their (individually) expected net gain is reduced in such a way that n_0 is the current criterion of newly arrived customers based on their present comparison of alternatives.

What is the optimal value, θ^*, or rather the optimal range, of the toll? Clearly this is given by

$$\frac{C}{\mu}(v_s-n_0-1)=R-\frac{C(n_0+1)}{\mu}<\theta^*\leqslant R-\frac{Cn_0}{\mu}=\frac{C}{\mu}(v_s-n_0). \tag{26}$$

If a toll taken from this range is levied on customers joining the queue, the combined income (in unit time) of customers and the revenue agency is maximized. We might explicitly mention that expenditure incurred in toll collection and in information processing is considered negligible in this presentation.

Clearly, if the toll revenue may be used for redistribution of income among the population or for socially useful purposes the proposed imposition of tolls is an optimal procedure.

6. REVENUE MAXIMIZATION

The toll collecting agency may be completely divorced from the individual and collective economic interests of the customers. In that case the agency will seek to impose a toll, θ_r, designed to maximize its own revenue rather than to optimize the whole system.

The objective function of the toll collector is given by

$$M=(\lambda-\zeta)\theta=\lambda\frac{1-\rho^n}{1-\rho^{n+1}}\left(R-\frac{Cn}{\mu}\right)=\lambda R\frac{1-\rho^n}{1-\rho^{n+1}}\left(1-\frac{n}{v_s}\right). \tag{27}$$

The maximization of M (which is considered a function of feasible $n-s$) is brought about by techniques similar to those used in previous sections. Let the appropriate value of n be designated by n_r. It is then possible to manipulate the inequalities associated with the maximum value of M in (27) in such a fashion that a convenient quantity v_r—analogous to v_0 in (23)—should be defined by

$$v_r+\frac{(1-\rho^{v_r-1})(1-\rho^{v_r+1})}{\rho^{v_r-1}(1-\rho)^2}=v_s. \tag{28}$$

The integer n_r which maximizes toll revenue is derived (as analogous integers before) by applying the bracket function on v_r:

[5] Such a measure would have to be explained very carefully to the participants since it is in apparent contradiction with "common sense."

(29) $n_r = [v_r]$.

Further (rather tedious) manipulation yields

(30) $v_r < v_0 < v_s$ (given $v_s > 1$),

the approximative meaning of which is the following. Some toll collection may be beneficial to a queueing system if an appropriate objective function (representing public good) is chosen. If the toll collecting agency is a decision maker tending to maximize its own revenue, however, the entrance fees θ_r, levied on joining customers, will be too high and social optimality will (frequently) not be attained:

(31) $$\theta_r = R - \frac{Cn_r}{\mu} = \frac{C}{\mu}(v_s - n_r) .$$

7. SOME CONCLUDING REMARKS

There is very little to add to the critique of the model and the general conclusions drawn from its structure. One point should be re-emphasized: The results in qualitative form are independent of the specifics of the model. Thus, for instance, if service times were distributed in some manner other then exponentially, we still would derive benefits from the collection of tolls, though the derivation of n_0 (or an equivalent doctrine) may be much more complex than that presented in this study.

The basic features of the model are shaped by the assumption of the existence of a public good and by the assumption of possible nonadmission of customers to the service station. Rewards are considered to be constant and equal. Again, no basically different results would have been obtained had these rewards been drawn from a distribution. A strong modification may be called for if we were to assume that the reward obtained depends in some way on the effective traffic density. Again, without going into detailed arguments, it can be shown that a policy of "laissez faire" is only rarely and accidentally a correct one (i.e., socially optimal). In this latter more general case, in which effective interaction between customers and therefore dependence on traffic density is assumed, the proper strategy is not necessarily the imposition of a toll; cases can be constructed where the handing out of subsidies to joining customers optimizes public good. The detailed analysis of such situations is the subject of further investigation.

Technion, Israel Institute of Technology,
and
University of North Carolina

REFERENCES

[1] LEEMAN, WAYNE A.: "The Reduction of Queues through the Use of Price," *Operations Research*, 12 (1964), 783–785.

[2] ———: "Comments" on Saaty's "The Burdens of Queueing Charges," *Operations Research*, 13 (1965), 680–681.

[3] SAATY, THOMAS L.: "The Burdens of Queueing Charges—Comment on a Letter by Leeman," *Operations Research*, 13 (1965), 679–680.