QUALITY, QUEUEING AND CAPACITY

Matthew F. Keblis
College of Business
University of Dallas
Irving, TX 75062-4736
Phone: (972) 265-5719
Fax: (972) 265-5750
Email: mkeblis@udallas.edu

ABSTRACT

We consider a model of customer's choice between companies producing a single good that varies only in price and quality. The quality (durability) of the good is specified exogenously while the price is determined endogenously. Customers can purchase one unit of the good from any of the companies. Each company serves customers using a single-server queueing model with a first-come-first-serve queue. The customer incurs a waiting cost per unit time. After receiving the good, the customer consumes it which takes a random time with a distribution determined by the quality of the good purchased. After consumption, the customer again purchases another unit of the good. It is assumed that the queues are not visible to the customers. The customers select companies so that their expected expenditure per unit time is minimized.

INTRODUCTION

We will consider how price and quality affect consumer choice. We assume a closed system of consumers each of who are in one of two states: waiting to purchase a good at a particular company or consuming the good that was recently purchased. There are a number of competing firms who sell similar goods that differ only in price and durability. The durability of goods is externally specified but the companies are free to choose their own prices in an attempt to attract more customers. The more customers that choose to purchase from a company the more the waiting line will grow. Consumers prefer low cost, long-lasting products but also prefer not to wait a long time in queue. A mixed strategy equilibrium will be found that determines the proportion of customers choosing each of the companies and will determine the prices that each company will charge.

THE MODEL

Consider n companies that produce an interchangeable product. The quality of the product of company k is measured by its lifetime before it needs replacement. This time is given by an exponential random variable with mean q_k. This mean is considered a value that the company has chosen, given its manufacturing capabilities and will never change. The company will

charge the consumer a price P_k per unit while having given fixed and variable costs of f_k and v_k respectively. The value of P_k is to be determined.

Initially, *N* consumers choose a company and purchase one unit of their good. Each company serves customers using a single-server queueing model with a first-come-first-serve queue (see Figure 1). The customer while waiting in the queue and while being served incurs a waiting cost per unit time w. A utility of ϕ per unit time is obtained while consuming the product. We assume for convenience that all companies have the same service capacity. Therefore the service time is assumed exponentially distributed with constant mean σ. After receiving the good, the customer consumes it during the random time with mean q_k. After consumption, the customer again purchases another unit of the good that can be chosen from any of the companies. It is assumed that the queues are not visible to the customers. The information available to the customer at this point is the current prices charged by each company and the given set of $\{q_k\}$. The customers choose a company so that their future discounted expected cost per unit time is minimized. Formally let $(p_1,...,p_n)$ be a probability vector representing a mixed strategy over the choice of company by a customer. Since the customers are assumed homogeneous we restrict consideration to a symmetric Nash equilibrium.

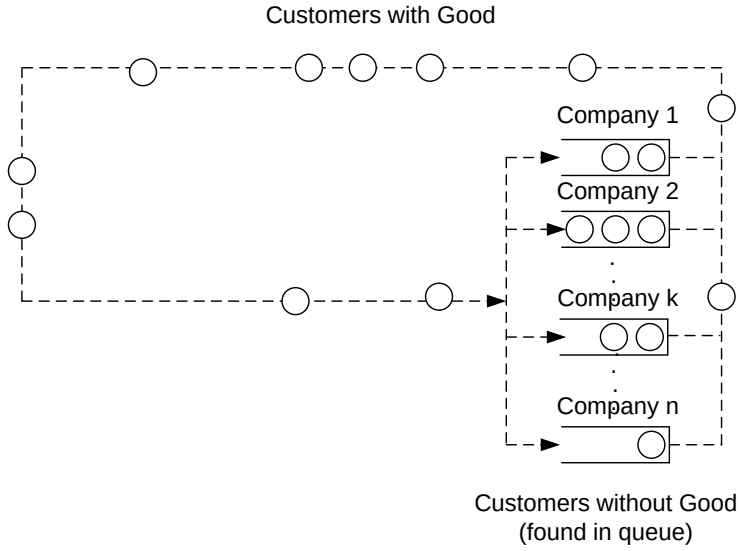

Figure 1

RELATED WORK

The earliest related work was undertaken by Naor (1969). He employed a queuing model with a cost structure to study the effect of tolls on customers. Both Luski (1976) and Levhari and Luski (1978) were among the first to model competition between service providers. In their problem settings customers chose their service provider on the basis of a full price where the full price had two pieces: a fixed piece and a piece that depended on the expected waiting time multiplied by a waiting cost rate. Loch (1991) considered a variant of the Luski (1976) model and was the

first to consider a problem setting with general servers, i.e. he went beyond studying the basic M/M/1 queueing model. Reitman (1991) generalized the results of Levhari and Luski (1978). Sattinger (2002) considered a special case of our model. He considered firms with identical quality goods. In equilibrium therefore the companies charged identical prices and customers were indifferent about which company to patronize. Chen and Wan (2003) also studied the problem settings of Luski (1976) and Levhari and Luski (1978) and showed that a price equilibrium exists for the basic model with a uniform cost rate. They showed however that the Nash equilibrium may fail to be unique. Christ and Avi-Itzhak (2002) showed that asymmetric Nash equilibrium of service rate pairs may arise when a customer arriving to empty queues is randomly assigned to a service provider. Armony and Haviv (2003) studied the situation where customers belong to one of two classes each with a different waiting cost parameter. The book by Hassin and Haviv (2003, chapter 7) contains an excellent overview and analysis of the previous work in this area. Most recently, Veeraraghavan and Debo (2009, 2011) showed how queue length can signal to a customer information about the quality of a service provider's offering. In the case where there is no waiting cost they found customers join the longest queue (Veeraraghavan and Debo, 2009). On the other hand when there are waiting costs, they found the equilibrium queue joining strategy is a complex function of both queue lengths (Veeraraghavan and Debo, 2011)

In this paper we consider an extension of the Sattinger model to heterogeneous companies while still assuming homogenous customers. The fact that we have a closed system with a fixed number of customers distinguishes it from all other models surveyed by Hassin and Haviv (2003). The outline of this paper is as follows. In the next section we use a Markov decision process to describe the decisions in the stochastic environment. The following two sections of the paper continue the analysis. The last section of the paper summarizes the results and outlines directions for future research.

THE MARKOV DECISION PROCESS

Start the decision making in the model at some arbitrary time 0. Suppose at some future time *t*, a customer has purchased a unit of the good and is currently consuming it. Let $V_G(t)$ be the expected discounted value over an infinite future of a customer having the good. Similarly, let $V_{NG}(t)$ be the expected discounted value over an infinite future, starting from time *t*, of a customer without the good and currently waiting in the queue (in equilibrium it will not matter which queue you are in.).

We assume that both the customers and the companies have used a given strategy for sufficient time so that the induced Markov process has reached a stationary distribution. And since queues are unobservable, the process will continue to evolve toward a stationary distribution. Furthermore, we assume this given strategy used by the participants is the Nash equilibrium. Therefore we can refer to constants V_G and V_G as the Nash equilibrium values of $V_G(t)$ and $V_{NG}(t)$ for large *t*. Furthermore, in the equilibrium, V_G will not depend upon the current queue *k*.

Define $\mu_k = 1/q_k$ as the failure rate of good *k*. Suppose, for convenience, we know that the flow of customers arrive at company *k* as a Poisson process with rate λ_k. Assume *r* is the rate of discounting so that a payoff *D* that will be obtained at a time *s* in the future has a present value of $e^{-rs}D$. If *s* is a small time interval then this becomes $[1-r\Delta t]D + o(\Delta t)$.

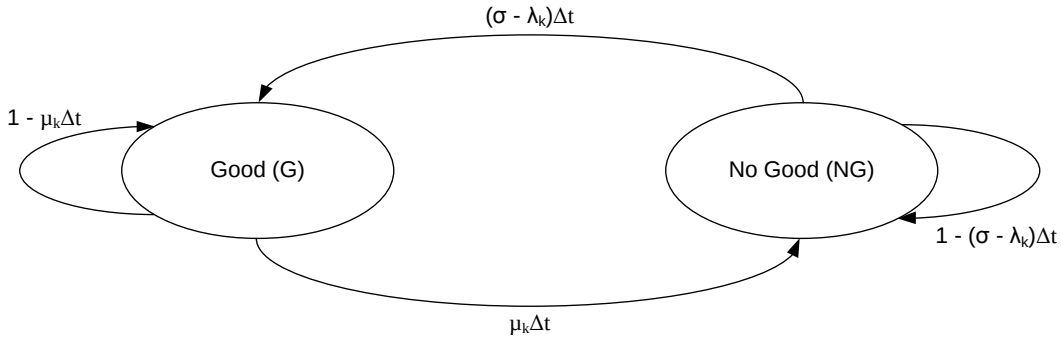

Figure 2

Now consider a two-state continuous time Markov process with states *G* and *NG* (see Figure 2). Assume we have a consumer who is a customer of company *k* and never purchases from another company. Customers alternate visits to the same M/M/1 queue with a consumption time. During a small time interval we will have either 0 or 1 transition. If in state *G*, the transition rate to state *NG* is μ_k. If in state *NG*, we assume the consumer is in the queue of company *k*. The waiting time in the system including completing service is exponential (Ross, 1993, p. 362) with mean $(\sigma - \lambda_k)^{-1}$. Therefore the transition matrix over a time interval Δt is, to within order $o(\Delta t)$:

$$\begin{array}{c} \\ G \\ NG \end{array} \begin{array}{c} \begin{array}{cc} G & NG \end{array} \\ \begin{bmatrix} 1-\mu_k\Delta t & \mu_k\Delta t \\ (\sigma-\lambda_k)\Delta t & 1-(\sigma-\lambda_k)\Delta t \end{bmatrix} \end{array} \tag{1}$$

If a customer is currently in state *G* at time *t*, then after time Δt, the customer will either transition to *NG* or stay in *G* (assume the transition is at the end of the time Δt). So in either case the customer will accumulate a utility of $\phi\Delta t$ from consumption while in state *G*. A customer's expected future payoff (payments received at time $t+\Delta t$) is $(1-\mu_k\Delta t)\left[V_G(t+\Delta t)+\phi\Delta t\right]+\mu_k\Delta t\left[V_{NG}(t+\Delta t)+\phi\Delta t\right]$. Multiplying this expression by $e^{-r\Delta t}$ will discount this back to the present time *t*. The result must equal $V_G(t)$, the value of being in the current state (we suppress a dependence on *t* in the V_G notation since the stationary case is assumed). Then we have: $V_G = (1-r\Delta t)\left\{(1-\mu_k\Delta t)\left[V_G+\phi\Delta t\right]+\mu_k\Delta t\left[V_{NG}+\phi\Delta t\right]\right\}+o(\Delta t)$

After some simplification we can divide by Δt and then let $\Delta t \to 0$. This yields the following relationship between V_G and V_{NG}:

$$rV_G = \phi + \mu_k(V_{NG} - V_G) \tag{2}$$

Similarly, if a customer is currently in state *NG* then at the end of the time interval Δt a customer might stay in *NG* or transition to state *G*. If a customer moves to *G*, then the customer

will pay P_k upon exiting the queueing system. In any case, the customer will incur a waiting cost of $w\Delta t$. After discounting, we obtain the expression

$$V_{NG} = (1 - r\Delta t)\left\{(1-(\sigma-\lambda_k)\Delta t)\left[V_{NG} - w\Delta t\right] + (\sigma-\lambda_k)\Delta t\left[V_G - w\Delta t - P_k\right]\right\} + o(\Delta t)$$

Proceeding as with the equation for V_G, this reduces to:

$$rV_{NG} = -w + (\sigma-\lambda_k)(V_G - V_{NG} - P_k) \tag{3}$$

From (2), $V_G = (\phi + \mu_k V_{NG})/(r+\mu_k)$. This can be substituted into (3) to obtain an expression for V_{NG}:

$$V_{NG} = \frac{(\sigma-\lambda_k)[\phi - P_k(r+\mu_k)] - w(r+\mu_k)}{r(r+\mu_k+\sigma-\lambda_k)} \tag{4}$$

From this we can obtain an equation for V_G:

$$V_G = \frac{\phi r(r+\mu_k+\sigma-\lambda_k) + \phi\mu_k(\sigma-\lambda_k) - \mu_k(r+\mu_k)[(\sigma-\lambda_k)P_k + w]}{r(r+\mu_k)(r+\mu_k+\sigma-\lambda_k)}$$

which can be simplified to

$$V_G = \frac{\phi}{r} - \frac{\mu_k[\phi + w + (\sigma-\lambda_k)P_k]}{r(r+\mu_k+\sigma-\lambda_k)} \tag{5}$$

PARALLEL QUEUES

At any given time let *G* be the total number of customers which have the good and *Q* be the number in a queue waiting to obtain the good. Then $N = G + Q$. Further, we can split *G* into the expected number holding each good: $G = \sum_{k=1}^{n} G_k$. Likewise, $Q = \sum_{k=1}^{n} Q_k$ where Q_k is the expected number in the queue *k* (including the one in service). The expected number in an M/M/1 queueing system is $Q_k = \lambda_k/(\sigma-\lambda_k)$. The process of consuming the good is a self-service or M/M/∞ queue. The output rate of queue *k* is λ_k which is thus the input rate for the self-service queue for those with good *k*. Therefore, $G_k = \lambda_k/\mu_k$. This assumes the $\{\lambda_k\}$ are fixed when, in fact, they actually are determined by the collective decisions of the customers.

Let $(p_1,...,p_n)$ be a probability distribution such that p_k is the probability that a customer chooses company *k*. We need to solve for $(p_1,...,p_n)$. In equilibrium, with *N* very large, we can view the system as *n* subsystems of *queues in parallel that do not interact*. Then there will be M_k customers in queue *k*, $M_k = Np_k$. Therefore,

$$M_k = G_k + Q_k = \lambda_k/\mu_k + \lambda_k/(\sigma-\lambda_k) \tag{6}$$

From (6), $M_k = \dfrac{\lambda_k(\sigma-\lambda_k) + \mu_k\lambda_k}{\mu_k(\sigma-\lambda_k)}$. This is a quadratic in λ_k which can be solved:

$$\lambda_k = 0.5[\sigma + (M_k+1)\mu_k] - 0.5\sqrt{[\sigma + (M_k+1)\mu_k]^2 - 4\sigma M_k\mu_k} \tag{7}$$

(We take the negative sign in the quadratic equation or otherwise $\lambda_k \geq \sigma$ and the queue length grows without bound.) This is equivalent to Sattinger's equation (11).

In our model we have *n* classes of customers since the queues are non-interacting. Each customer class will determine a V_{NG} value given by (4). Each customer will try to maximize this value (equivalently V_G; one determines the other). This will be accomplished by changing the strategy $(p_1,...,p_n)$. The result is that each company will be equally attractive in the sense that all the V_{NG} values are the same (not that customers choose each company with equal probability). So denote this common value of V_{NG} for each queue as *V*. Now we can use (4) to solve for the arrival rate to queue *k*:

$$\lambda_k = \sigma - \frac{(w+rV)(r+\mu_k)}{\phi - P_k(r+\mu_k) - rV} \tag{8}$$

This is the same as Sattinger's equation (12). The denominator must be positive so this places an upper bound on possible values of *V*, namely: $V < [\phi - (r+\mu_k)P_k]/r$.

For any given *V* we can find the required $\{\lambda_k\}$ from (8). Using them in (6) we obtain:

$$p_k = \frac{\lambda_k(\sigma - \lambda_k) + \mu_k\lambda_k}{N\mu_k(\sigma - \lambda_k)} \tag{9}$$

From this result we require $\sum_{k=1}^{n} p_k = 1$. This places a restriction on the $\{\lambda_k\}$ which means *V* can be determined in this manner (although there may not always be a solution).

This is different from Sattinger's model since we are now assuming a large but finite N. This is only an approximation since this assumption conflicts with the statement that all *n* queues operate independently.

THE TANGENT LINE

In equation (8), fix the customer's utility at *V* and find $dP_k/d\lambda_k$ along the constant utility curve. Taking the derivative of each side of (8) yields:

$$\frac{dP_k}{d\lambda_k} = -\frac{[\phi - (r+\mu_k)P_k - rV]^2}{(w+rV)(r+\mu_k)^2} \tag{10}$$

Now the company profit function is $\Pi_k = (P_k - v_k)\lambda_k - f_k$. Fix this profit for various (P_k,λ_k) values and find $dP_k/d\lambda_k$ along this isoprofit curve. Then $\frac{\partial \Pi_k}{\partial \lambda_k} = P_k + \lambda_k \frac{dP_k}{d\lambda_k} - v_k = 0$ and so $dP_k/d\lambda_k = -(P_k - v_k)/\lambda_k$. At the equilibrium point the $dP_k/d\lambda_k$ for the customer and for the company must be equal. This yields the relation:

$$P_k = v_k + \frac{\lambda_k[\phi - (r+\mu_k)P_k - rV]^2}{(w+rV)(r+\mu_k)^2} \tag{11}$$

We can now substitute the value of *V* from (4) into (11) to give us P_k as a function only of λ_k, plus constants:

$$P_k = v_k + \lambda_k \frac{(\phi + w) - v_k(r + \mu_k)}{(\sigma - \lambda_k)^2 + \sigma(r + \mu_k)} \tag{12}$$

This is Sattinger's (13).

As N increases, the value of V (same for all queues) decreases since everything is getting more congested. Eventually N will get so large that $V = 0$. This is the maximum capacity of the system. (We could put more customers into the system but then $V < 0$ and they would prefer, if possible, to exit the system.)

To find the characteristics of the system at the maximum capacity, set $V = 0$ in (8) and (11) and we find that $\lambda_k = \sigma - \frac{w(r + \mu_k)}{\phi - P_k(r + \mu_k)}$ and $P_k = v_k + \frac{\lambda_k[\phi - (r + \mu_k)P_k]^2}{w(r + \mu_k)^2}$.
Substitute the first equation into the second to obtain, after some algebra,

$$P_k = \frac{\phi}{r + \mu_k} - \sqrt{\frac{w[\phi - v_k(r + \mu_k)]}{\sigma(r + \mu_k)}} \tag{13}$$

and

$$\lambda_k = \sigma - \sqrt{\frac{\sigma w(r + \mu_k)}{\phi - v_k(r + \mu_k)}} \tag{14}$$

at the maximum capacity. The corresponding $\{p_k\}$ now can be found from (9). This will just say that the $\{p_k\}$ are proportional to the number in each "parallel system."

SUMMARY AND FUTURE RESEARCH

In paper we consider a model of customer's choice between companies producing a single good that varies only in price and quality. We assume the quality (durability) of the good is specified exogenously while the price is determined endogenously. A mixed strategy equilibrium is found that determines the proportion of customers choosing each of the companies and the prices that the companies will charge. Additional work on this problem will involve undertaking a numerical study that will lead to a comparison with the results found in Sattinger (2002). Furthermore, future research needs to investigate the “small N” case, i.e. where the number of customers is few in number. It would also be interesting to relax the assumption that the queues are not visible to customers as very little research has been done on this front (see Hassin and Haviv (2003, chapter 8)).

REFERENCES

Armony, M. and Haviv, M. (2003). Price and delay competition between two service providers, *European Journal of Operations Research* 147(1), 32-50.

Chen, H. and Wan, Y.-W. (2003). Price competition of make-to-order firms, *IIE Transactions* 35(9), 817-832.

Christ, D. and Avi-Itzhak, B. (2002). Strategic equilibrium for a pair of competing servers with convex cost and balking, *Management Science* 48(6), 813-820.

Hassin, R. and Haviv, M. (2003). To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems, Kluwer Academic Publishers: Boston.

Levhari, D. and Luski, I. (1978). Duopoly pricing and waiting lines, *European Economic Review* 11, 17-35.

Loch, C. (1991). Pricing in markets sensitive to delay. Ph.D. Dissertation, Stanford University.

Luski, I. (1976). Partial equilibrium in a queueing system with 2 servers, *Review of Economic Studies* 43(3), 519-525.

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

Reitman, D. (1991). Endogenous quality differentiation in congested markets, *Journal of Industrial Economics* 39(6), 621-647.

Sattinger, M. (2002). A queuing model of the market for access to trading partners, *International Economic Review* 43(2), 533-547.

Veeraraghavan, S. and Debo, L. (2009). Joining longer queues: Information externalities in queue choice, *Manufacturing and Service Operations Management* 11(4), 543-562.

Veeraraghavan, S. and Debo, L. (2011). Herding in queues with waiting costs: Rationality and regret, *Manufacturing and Service Operations Management* 13(3), 329-346.