To Join the Shortest or the Longest Queue?  
Inferring Service Quality from Queue Lengths.

A classical example that illustrates how observed customer behavior impacts other customers’ decisions is the selection of a restaurant whose quality is unsure: Customers often choose a busier restaurant despite additional waiting time they have to incur, inferring that other customers in that restaurant know something that they do not know. When choosing between different service providers of which the quality is common knowledge, customers trade off differences in expected waiting times for services with differences in service value when deciding from which firm to purchase the service. In the case that the quality of the service provider is not common knowledge, but customers have some private information about the service quality, observed behavior of other customers may complement one’s private information. In an environment where congestion occurs, queues are an indication of the other customers’ choices. Rational customers take thus congestion levels or queue lengths at two competing service providers into account when making a purchasing decision. It may even be the case that customers ignore their private information and make a decision solely based on the observed queue lengths. In the economics literature, this has been referred to as ‘herding’. In this paper, we introduce herding behavior of customers, in a two-server queuing model. By means of a simple model, we study how two service providers split the market. Customers arrive according to a Poisson process to the market. Service times are exponentially distributed with the same rate at each service provider. Upon arrival customers receive some private, but, imperfect information about which server provides better quality. In addition, customers observe the number of other customers queuing at each server. Based on both private and queue length information customers decide which queue to join or whether to balk. We find that it may be rational in equilibrium to ignore private information and purchase from the service provider with the longer queue. Our model allows better understanding of how the queue length impacts consumer choice for products or services with quality uncertainty.

1. **Introduction**

In many real life situations, we must choose between alternative providers of a certain service of which the value is not perfectly known in advance. Examples are selecting a restaurant, a web service, a movie, attending a sports event etc. In general, customers rely on publicly available information, but, may also have some private information guiding which service provider to choose. The examples cited above all have in common that they create congestion; waiting lines in front of a restaurant, delays for a web service or queues for a movie or a sports event. It is not unreasonable to expect that customers also take congestion levels
into account when making their selection: High levels of congestion at one server, but, not at the other one may be an indication that the customers that chose that one server did so because of their private information of its superior value compared to the other server. In other words: Long queues may be an indication of high service quality. On the other hand, long queues may decrease the prospective customer’s net utility due to high waiting costs. As congestion is also generated by randomness on either the customer arrival process or the service provision process, long queues may be created by chance, possibly triggering other customers to join the same queue. We expect thus that dynamics play an important role: It may take a long time for a the least popular service provider (with the shortest queue) to attract a critical number of customers despite offering the best quality in the market.

While the problem description above seems to be quite generic and different literatures have studied aspects of it, to the best of our knowledge, no study has incorporated how queues impact customer purchasing behavior when the service quality is uncertain. This paper is first attempt at understanding this issue through a simple two server queueing system that captures key characteristics of the generic situation explained above: Service providers are equal ex ante and have an infinite waiting space. Customers arrive according to a Poisson Process, observe the queue lengths at both servers and some private information about which server provides the best quality. Based on this information, they decide which queue to join or whether to balk. Customers are served in a first come first served basis. There is no jockeying or reneging in the system. We assume that all customers are rational Bayesian decision makers that maximize their expected utility. Our model allows us to address the following questions: When do customers ignore their private information and make their purchasing decision based on the observed queue length information? How does the signalling function of queues impact the formation and decay of these queues?

2. Literature Review

Based on our generic description, we are interested in learning more about (1) how different agents in a market influence each other’s purchasing decisions when the quality of a good or service is uncertain and (2) how congestion impacts the service provider selection. We found related economics, operations research and operations management literature that we discussed briefly below. None of these papers, however, provide insight in to how both the phenomena interact simultaneously.
Economics literature: Becker (1991) first noted that a popular seafood restaurant (in Palo Alto) has every day long queues during every day while the restaurant across the street with comparable food has many empty seats most of the time. Becker explains why firms even with persistent excess demand do not raise prices by assuming that demand for the good by a person depends on the aggregate quantity demanded in the market. Becker focuses on monopoly pricing and does not consider information asymmetry. Bikhchandani, Hirshleifer and Welch (BHW 1992) explicitly model the role that observed behavior of other actors plays when purchasing an asset with not perfectly know value. They explain how informational cascades are created. Informational cascades occur when a series of actors take a decision based on private information that is observed by others and then every subsequent actor, based on the observations of others, makes the same choice independent of his/her private signal. While BHW explicitly model the information structure, they do not incorporate any supply effects like congestion or crowding that may occur when all agents make the same choice. Chamley (2004) provides a comprehensive survey on herding literature. Hassin and Raviv (2002) provide a detailed analysis on economics of queuing theory.

In queueing theory, there exists a large literature that focusses deciding which queue to join. Whinston (1977) showed that in a purely Markovian model with Poisson arrival processes and exponentially distributed service times that are independent, joining the shortest queue upon arrival does not only minimize expected individual delay, but also the long-run average delay per customer. Whitt (1986) provides examples of service time distributions that are not exponential for which choosing the shortest queue does not necessarily minimize the long-run average delay per customer. In this literature, there is no uncertainty on the expected service quality at each server. Gans (2000) studies customer loyalty to a supplier by explicitly modeling uncertainty on the expected service quality of a set of service providers and allows customers to learn the true quality of every service provider through (expensive) repeated service sampling. A single customer has to determine at each trial which service provider to visit. While Gans explicitly studies the uncertainty on the service quality, he focuses on a single customer that repeatedly interacts with one or more suppliers.

3. The Model

The game: We consider a game in which customers arrive sequentially according to a Poisson process with arrival rate $\lambda$ to a market with two servers, labeled by $i = 1$ and
i = 2. Customers are labeled by the order at which they arrive, t ∈ {1, 2, . . .}. Upon his arrival, customer t observes the number of other customers queueing at both service providers, n = (n₁, n₂). Each customer decides whether to purchase the service or not. In the latter case, the customer ‘balks’. In the former case, the customer decides from which service provider to purchase the service (i.e. which queue to join).

The service value: Service time at both servers is exponentially distributed with mean $\frac{1}{\mu}$. The exact service value is unknown to the customers, but, its joint distribution, $F(v₁, v₂)$ is common knowledge (with $f(v₁, v₂)$ the density function). In addition to observing the queue length at both servers, each customer receives a private signal, $s$. The signal is an imperfect indicator of which server provides the highest value; $s$ is such that: $\Pr(s = 1 \mid V₁ > V₂) = \Pr(s = 2 \mid V₁ < V₂) = g$, i.e. if the true state is that server $i$ provides better value than server $j$, each agent receives signal $s = i$ ($s = j$) with probability $g \ (1 - g)$, where $g > \frac{1}{2}$.

Consistent Customer Behavior: Consider the $t^{th}$ customer to arrive at the market. Let $S = \{1, 2\}$ and $T(t) = \{0, \ldots, t - 1\}$. A mixed strategy for this customer is then a mapping $\sigma_t : S \times T(t) \rightarrow [0, 1] \times [0, 1]$. Let $\sigma_t(i, s, n)$ be the probability that customer $t$ joins queue $i$ after observing signal $s$ and state $n$. Finally, let $\Sigma_t = \{\sigma_1, \ldots, \sigma_t\}$ denote the combined strategy profile of all customers $1, \ldots, t$. By convention, we denote $\Sigma_0 = \emptyset$. The random arrival and service processes, in conjunction with the strategies of customers $1, \ldots, t - 1$, induce a probability distribution over $T(t)$. Let $\pi_t(i, n; \Sigma_{t-1})$ denote the probability that there are exactly $n$ people in the queues when the $t^{th}$ customer arrives, conditional on $V_i > V_{-i}$, with $-i$ denoting $2(1)$ if $i = 1(2)$ and conditional on the combined strategy profile of all previous customers being $\Sigma_{t-1}$.

After observing $(n, s)$, the customer $t$ can update the (common) prior on the service values, $f(v₁, v₂)$ to $f_t(v₁, v₂|n, s, \Sigma_t)$, and update the service value as follows:

$$E(V_t(i) \mid n, s; \Sigma_t) = \int_0^\infty \int_0^\infty v_i f_t(v₁, v₂ \mid n, s; \Sigma_t) dv₂ dv₁, \ i, s \in \{1, 2\}$$

Rational Customer Actions: Assume that $\pi_t(i, s, n; \Sigma_{t-1})$ are given, then, we can determine the optimal policy of the customer observing $(n, s)$: Let $a_t(s, n; \Sigma_{t-1}) = 1(2)$ if the customer $t$ joins server $1(2)$, and 0 otherwise. The optimal consumer decisions are determined by:

$$E(V_t(s) \mid n, s; \Sigma_{t-1}) - cn_s/\mu > [E(V_t(s) \mid n, s; \Sigma_{t-1}) - cn_s/\mu]^+ \Rightarrow a_t^*(s, n; \Sigma_{t-1}) = s$$
Table 1: Possible strategies as a function of the private signal. The $\cdot$ represents $n, l$.

<table>
<thead>
<tr>
<th>$a(s, \cdot)$</th>
<th>$a(1, \cdot) = 0$</th>
<th>$a(1, \cdot) = 1$</th>
<th>$a(1, \cdot) = 2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a(2, \cdot) = 0$</td>
<td>Always Balk</td>
<td>Follow Signal 1 Balk Signal 2</td>
<td>Contradict Signal 1 Balk Signal 2</td>
</tr>
<tr>
<td>$a(2, \cdot) = 1$</td>
<td>Contradict Signal 2 Balk Signal 1</td>
<td>Always Join Queue 1</td>
<td>Contradict Signals</td>
</tr>
<tr>
<td>$a(2, \cdot) = 2$</td>
<td>Follow Signal 2 Balk Signal 1</td>
<td>Follow Signals</td>
<td>Always Join Queue 2</td>
</tr>
</tbody>
</table>

and

$$
\max_i E(V_i(i) \mid n, s; \Sigma_{t-1}) - cn_i/\mu < 0 \Rightarrow a_t^*(s, n; \Sigma_{t-1}) = 0
$$

A Bayesian equilibrium is determined by rational customer actions, $\sigma^*_t$ with beliefs $f_t(v_1, v_2|n, s, \Sigma^*_{t-1})$ that are consistent with Bayes’ Theorem. Note that $a_t^*(s, n; \Sigma_{t-1})$ from equation (??) determines $\sigma^*_t$ and thus $\Sigma^*_t = \Sigma^*_{t-1} \cup \{\sigma^*_t\}$.

We can derive conditions for a Bayesian equilibrium when the prior distribution of the service values is uniform over $[0, v]$ and are independent from each other. Table 1 represents all possible strategies as a function of the private signal.

We can determine rational strategies as a function of the state of the queue, $n$ and the likelihood ratio, $l_t(s, n; \Sigma_{t-1}) = \frac{\pi_t(1,s,n;\Sigma_{t-1})}{\pi_t(2,s,n;\Sigma_{t-1})}$. It follows thus that there are six non-trivial strategies to consider, see Table ??: (i) ignore the private information and join always queue 1, or queue 2, or balk, (ii) follow the private information, (iii) follow one signal and balk otherwise. Intuitively, doing exactly the opposite to both signals, or, doing the opposite to one signal and balk for the other signal are never rational, as $g > \frac{1}{2}$.

4. Insights and Conclusions

There has been little work done on consumer purchase behavior when presented with service or quality uncertainty, and information about congestion. We analyze competing service providers with queues with waiting costs and without waiting costs. In our context, we define herding as consumers joining the longer queue contrary to one’s own private signal.

When there is no waiting cost, following the longer queue and following the signal when queues are equally long, is an equilibrium strategy. We term this as the “ladder” strategy. Under this equilibrium strategy, one of the two service centers is always empty, but, which one is empty may change over time. Within the class of “bandwidth” strategies (namely,
the strategies that compare the differences in queue lengths), the single bandwidth strategy is also an equilibrium strategy. This bandwidth strategy is to follow the signal when the queue lengths differ by one (or less), and otherwise to follow the longer queue. The ladder and bandwidth strategies reflect herding behavior, as the queues are purely informative and do not impose any negative externality (i.e. waiting costs).

For queues with positive waiting costs, we find always following one’s signal is never an equilibrium strategy. Hence the decision whether to heed one’s private signal is highly dependent on the congestion levels in the system. The model becomes intricate even for simple bandwidth strategies. Under certain restrictions, we can define modified bandwidth strategies. There exist threshold levels where following the shorter queue is better, and lower thresholds where it is attractive to follow the longer queue.

What we have considered, is a foray into the issue of understanding how the dynamics of congestion (positively or negatively) influence customer choice when quality uncertainty exists. We study, how the herding effects plays against the waiting costs and private information.

References


