M&SOM MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Vol. 11, No. 4, Fall 2009, pp. 543–562
ISSN 1523-4614 | EISSN 1526-5498 | 09 | 1104 | 0543

informs®
DOI 10.1287/msom.1080.0239

Joining Longer Queues: Information Externalities in Queue Choice

Senthil Veeraraghavan
The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104, senthilv@wharton.upenn.edu

Laurens Debo
Graduate School of Business, University of Chicago, Chicago, Illinois 60637, ldebo@chicagogsb.edu

A classic example that illustrates how observed customer behavior impacts other customers' decisions is the selection of a restaurant whose quality is uncertain. Customers often choose the busier restaurant, inferring that other customers in that restaurant know something that they do not. In an environment with random arrival and service times, customer behavior is reflected in the lengths of the queues that form at the individual servers. Therefore, queue lengths could signal two factors—potentially higher arrivals to the server or potentially slower service at the server. In this paper, we focus on both factors when customers' waiting costs are negligible. This allows us to understand how information externalities due to congestion impact customers' service choice behavior.

In our model, based on private information about both the service-quality and queue-length information, customers decide which queue to join. When the service rates are the same and known, we confirm that it may be rational to ignore private information and purchase from the service provider with the longer queue when only one additional customer is present in the longer queue. We find that, due to the information externalities contained in queue lengths, there exist cycles during which one service firm is thriving whereas the other is not. Which service provider is thriving depends on luck; i.e., it is determined by the private signal of the customer arriving when both service providers are idle. These phenomena continue to hold when each service facility has multiple servers, or when a facility may go out of business when it cannot attract customers for a certain amount of time. Finally, we find that when the service rates are unknown but are negatively correlated with service values, our results are strengthened; long queues are now doubly informative. The market share of the high-quality firm is higher when there is service rate uncertainty, and it increases as the service rate decreases. When the service rates are positively correlated with unknown service values, long queues become less informative and customers might even join shorter queues.

Key words: customer herding; behavioral operations; information externalities; market dynamics; private signals; service capacity
History: Received: April 24, 2007; accepted: June 26, 2008. Published online in *Articles in Advance* October 24, 2008.

1. Introduction

In many real-life situations, we often have to choose between service providers whose quality is not perfectly known in advance. For example, when selecting a restaurant at which to dine, a movie to watch, or a sports event to attend, we often do not know the exact valuation of the experience (although we might have an expectation about the service). The examples above have the following common feature: They create congestion. Waiting lines in front of a restaurant, or queues for a movie or a sports event are commonly observed. Therefore, it is not unreasonable to expect that we are influenced by the "level of business" or "buzz" at each service facility when making our selection. In other words, we complement our incomplete private information regarding which service provider to choose with publicly available information such as observed congestion at one service facility.

Long queues at one server but not at another may provide an indication that several customers chose that particular service, perhaps because of its perceived superior value compared to the other server. Because queues are typically generated by randomness of 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. What then, is the information contained in a queue that forms in front of a service? A long queue in front of a service provider might be an indication of slow service. On the other hand, a long queue in front of the service provider might be an indication that it provides higher quality of service and therefore more customers have chosen that service. However, if more and more people join a longer queue, ignoring their own private valuation, then the signal quality provided by the queue weakens. Therefore, it is unclear what information a longer queue might contain.

Consider a customer who has the flexibility to watch a touring Broadway show that is playing in town. She has to decide on only one of two shows X and Y (perhaps due to limited budget). Because many such shows are experience goods, she does not surely know which show is better. She also does not mind seeing a show a few days later if tickets are unavailable immediately. Her private belief is that X might be a better show, but like every other customer, she is unsure about it. Suppose that the customer could get tickets for show X almost immediately, whereas the show Y is not available for, say, 10 days. Hence, she notices that Y is more popular than X. She can infer that one show is considerably less popular than the other by using the waiting information of 10 days. However, she also knows that other customers would have made different observations and their own rational decisions. She wonders how those choices of other customers would have in turn influenced what she sees. Using all this information, she determines which show is better and makes her choice. At an aggregate level, the choices of many customers will be linked with each other and determine the aggregate demand that each show attracts.

In a famous example, Becker (1991, p. 1110) notes a puzzling implication of how individual decisions that are linked with each other impact the aggregate demand of two seafood restaurants in Palo Alto; one is always crowded and the other nearly empty, even though they have similar prices and amenities, and writes,

> Suppose that the pleasure from a good is greater when many people want to consume it, perhaps because a person does not wish to be out of step with what is popular or because confidence in the quality of food, writing, or performance is greater when the restaurant, book, or theater is more popular.

Such an assumption would indicate that all customers have a higher ex post utility when they consume the more popular product. Becker's model is static and postulates that demand for a good is directly and positively influenced by other consumers' demand or its popularity by assuming a functional form.

Motivated by the examples mentioned, we posit that customers do not make a service selection decision in a "vacuum," but they are influenced by what they observe around them and, in particular, by the congestion at the service providers. As queue dynamics play an important role in the service context, we develop a simple, stylized, two-server queueing system that captures the aforementioned idea. These features have not yet been explored jointly in the previous literature. In our model, service providers are identical ex ante and have an infinite waiting space. By allowing the buffer space to be infinite, we can focus on the information contained by queues. The customers, each carrying some private information about which server provides the best quality, arrive according to a Poisson process and observe the queue lengths at both of the servers. Based on this information, they decide which queue to join. There is no jockeying or reneging in the system. We assume that all customers are rational Bayesian decision makers that maximize their expected utility.

The dynamics introduced in such systems may be intricate. Due to uncertainty about the service value, it may be possible that customers purchase the service that provides lower value. This may even result in an inferior service provider becoming more popular, and hence a longer queue forms at the inferior service provider. It may take a long time for the less popular service provider (the one with the shortest queue) to attract customers despite offering the best quality in the market.

Our model allows us to answer the following questions: When do customers ignore their private information and make their purchasing decision based on the observed queue-length information? How do service rates of service providers affect the queue-choice behavior? How does information from the queue lengths impact the formation and decay of these queues? How long do services remain popular (i.e., have long queues)? Could a server that provides higher-valuation service perform poorly against another inferior service provider with the same service rate? Could a slower service provider gain a higher market share by providing better service?

The paper starts by reviewing the related literature in §2, and then outlining our analytical model in §3. In §4, we develop the conditions for pure-strategy equilibria to exist, and then discuss the properties of resulting equilibrium strategies for service providers with symmetric service rates. Focusing initially on the equal service rates allows us to study the effect of arrival rates alone on customer choice behavior. We analyze various extensions of the symmetric service model, including multiserver queues and service providers' survival in equilibrium in §5. We then study the effect of service departures on customer choice behavior in §6 by modeling asymmetric service rates. Finally, in the concluding §7, we discuss our insights and future research directions.

2. Related Literature

We are interested in studying how different agents in a market influence each other's purchasing decisions when the quality of a good or service is uncertain and there is potential for congestion to impact service provider selection. Related issues have been addressed separately in the economics, queue choice, and operations management literatures, but the interaction between congestion, service rate, and service choice between service providers has not been studied. In the following paragraphs, we describe the literature in each area and differentiate our model.

2.1. Economics Literature

Becker (1991) observed that a popular seafood restaurant (in Palo Alto) had long queues during prime hours every day, whereas the restaurant across the street, with comparable food, had many empty seats most of the time. In explaining the phenomenon, he focuses on monopoly pricing and perfect information. Becker explains why consumer demand is very "fickle" and why "shift of restaurants between 'in' and 'out' categories occurs" using a pair of equilibria that are unstable under large demand changes. Chamley (2004) notes that an analysis of the above interesting problem that employs "optimization behavior for the consumers, and a dynamic analysis with imperfect information…" remains to be done. We examine how the observation made by Becker (1991) can be explained by modeling the rational behavior of customers in a queueing system.

Bikhchandani et al. (BHW, 1992) study the role that the observed behavior of other actors plays when purchasing an asset whose value is not perfectly known. They explain how informational cascades are created. Informational cascades occur when a series of actors makes a decision that is observed by every subsequent actor, each of whom also makes the same decision independent of his/her private signal. BHW assume that the entire history of the agents' decisions and their sequence is available to every arriving agent. BHW do not incorporate any supply effects such as congestion or departure of observed customers from the system due to completion of service. BHW find that a commonly observed history of actions can dominate private beliefs. In such cases, agents will choose actions that will not reveal their private information.

However, in many cases, arriving customers may not observe the entire history. For instance, nothing might be known about the decisions of customers whose service had been completed. Therefore, the information set of customers in our paper modeled along the following ways. First, the customer does not know his arrival sequence number. Second, the arriving customer does not know the full history of prior decisions. Finally, every arriving customer sees only the number of current customers and their queue choices, but does not see the order those customers arrived.

To our knowledge, there are only a few papers in herding literature that consider herding behavior under a limited amount of information. Smith and Sorensen (1998) consider a model where all arriving agents sample exactly two observations. The actions that each agent observes are exogenously generated using a "seed" population. Smith and Sorensen are interested in the probability of convergence to the truth. In Banerjee and Fudenberg (2004), in each period a continuum of customers simultaneously chooses their actions after observing exogenously chosen N previous actions. The aforementioned papers do not model endogenous censoring of information, and are concerned about convergence of public belief. Our paper differs from the analysis in both papers in several ways. First, we do not restrict all the customers to see the same number of observations. Moreover, the probabilities of those observations are endogenously generated in our model. Finally, ours is a stationary model.

In this respect, the closest paper to ours is a working paper by Callander and Horner (2006), who consider a queuing system with a restricted state space. Callander and Horner focus on market heterogeneity, i.e., how the agents in the market are differentially informed, and argue how a minority of informed agents can cause other uninformed agents to follow shorter queues. In contrast, we study a market where all customers are equally informed, and focus on understanding the queue lengths at which herding occurs. In addition, we also study the effect of service rates that are correlated with service value on customer choice behavior. In contrast with herding literature, depending on the service rates of the service providers or strength of the signal, we find that customers might *not* herd at all.

2.2. Queue Choice and Operations Management Literature

There is a wealth of literature that discusses the decision process of choosing between queues. Hassin and Haviv (2003) provide a comprehensive survey of the literature in equilibrium behavior of customers and service providers in queuing systems. The effect of negative externalities on choice between two queues has been well studied (Whinston 1977, Whitt 1986). Although the focus has been on joining shorter queues, simple characterizations of the probability distributions do not exist even for join-the-shortest queue discipline. The steady-state probabilities generated by joining the longer queues have not been studied, perhaps because of perceived lack of applicability.

Consequently, the number of papers that model positive externalities in multiple queues is limited. In a working paper by Debo et al. (2007), customers arriving in the market choose between joining a single queue or not. When there are no waiting costs, they find a threshold length below which customers with pessimistic information do not join and above which they do join. When there are waiting costs, they show that a nonthreshold strategy may determine the equilibrium. In contrast, using a richer two-queue model, we are able to show that both positive and negative externalities in queues could arise due to queue-length information alone.

Su and Zenios (2004, 2005) study patient choice in kidney allocation in a queueing context. Gans (2002) studies customers who choose among various service providers with uncertain quality. Customers learn the true quality of every service provider through (expensive) repeated service sampling. There are no congestion externalities in the model (i.e., each customer learns about the service only by experiencing the service and not by observing the choices of other customers). At each service episode a single customer determines which service provider to visit. Lariviere and van Mieghem (2004) model a system in which customers find congestion costly, and therefore plan to arrive when the system is underutilized. They show that when customers choose arrival times strategically, the equilibrium arrival pattern approaches a Poisson process as the number of customers gets large. In our paper, customers arriving according to a Poisson process have the option to choose from different service providers after having observed the queue length at each service provider. We show that if the service providers are symmetric in service rates, then longest-queue joining behavior occurs as soon as the queue difference is one.

Our model differs from existing literature in several aspects. We focus on information externalities and show that uncertainty in service valuations alone may either cause customers to join longer queues or to avoid longer queues. We demonstrate that if the service rates are different enough, customers might ignore the queue lengths completely and may follow their imperfect private signals. If the service valuations are positively correlated with speed, we find that customers may not necessarily join longer (or slower) queues. We show that in contrast to the existing literature, providers of a better-quality service in a market containing uninformed customers might have a higher market share by having lower service capacity (lower service speed). Thus, investing in promotion efforts to indicate quality could be more attractive than investing in higher service speeds or increasing service value further.

3. Model

In this section, we discuss the players, information sets, and decisions of the customers in the game we model. Next, we determine the conditions for equilibrium. These conditions allow us to find an equilibrium whose qualitative properties are discussed in the conclusion.

3.1. Market Arrivals and Service Rates

We consider a game in which customers arrive sequentially according to a Poisson process with arrival rate λ to a market with two servers, labeled $i=1$ and $i=2$. We assume that the service times at both servers are exponentially distributed with mean $1/\mu$. We allow the servers to have different service rates in §6. Also define $\rho=\lambda/\mu$. We assume that $\rho<1$, allowing a single server to capture the market. The service discipline does not need to be first-come-first-served (FCFS), as long as the service rates are exponentially distributed.

3.2. Service Value, Public Observation, and Private Signals

The exact service value of the service provided by a server i, v_i, is unknown to the customers. We assume that $(v_1, v_2) \in \{(v_h, v_l), (v_l, v_h)\}$, where $v_h > v_l > 0$. Thus, if $v_i = v_h$, then $v_{-i} = v_l$, where $i = 1, 2$ and $-i = 2, 1$. The initial priors are symmetric; i.e., $\Pr[v_1 = v_h] = \Pr[v_1 = v_l] = 1/2$. Without loss of generality, we can assume that service values are fixed, and that server 1 is better. Upon arrival at the market, customers observe the queue length in front of each service provider; $\mathbf{n} = (n_1, n_2)$. We refer to $\mathbf{n}$ as the state of the system. We assume that the customers incur *no* waiting costs. This assumption helps us focus on the information value contained in the length of the queue. Our results are sensitive to the presence of waiting costs, and we point out the complexity of analyzing waiting costs in §5. Following the assumption, we note that the value eventually accrued from the service is either v_h or v_l, regardless of the length of the queue. However, the expected value of the service is updated with the queue-length information. In addition to observing the queue length at both servers, each customer receives a private signal, $s \in \mathcal{S} = \{1, 2\}$, where $\mathcal{S}$ is the set of private signals. The signal is an imperfect indicator of which server provides the highest value; $s \in \mathcal{S}$ is such that $\Pr[s=1 \mid v_1 > v_2] = \Pr[s = 2 \mid v_1 < v_2] = g$; i.e., if the true state of nature is that server i provides better value than server j, each agent receives signal $s=i$ with probability g and signal $s=j$ with probability $1-g$. Note that the signal is imperfect because $1/2 < g < 1$. When $g=1$, the signal completely reveals the better service provider, and every customer would then choose the better server, ignoring the choices of other customers. When $g = 1/2$, the signal does not reveal any information about the quality of the service providers. We term such a signal as being completely noisy.

3.3. Customer Behavior

Consider any customer to arrive at the market. Let $\mathcal{A} = \{1, 2\}$ be the set of possible actions that the customer can take upon arrival; 1 represents joining server 1, and 2 represents joining server 2. A mixed strategy for this customer is then a mapping $\sigma\colon \mathcal{A} \times \mathcal{S} \times \mathbb{N}^2 \to [0, 1]$. Let $\sigma_j(a, s, \mathbf{n})$ be the probability that this customer j joins queue a after observing signal s and state $\mathbf{n}$ (with $\sum_{a\in\mathcal{A}} \sigma_j(a, s, \mathbf{n}) = 1$). As all customers are homogeneous ex ante, we consider only symmetric strategies. For a given strategy σ, let $\pi_i(\mathbf{n}, \sigma)$ be the long-run probability that the system state is $\mathbf{n}$ conditional on $v_i > v_{-i}$, with $-i$ denoting 2 (1) if $i=1$ (2), with σ being the customer's strategy. With the PASTA property (Wolff 1982), $\pi_i(\mathbf{n}, \sigma)$ is also the probability that the system is in state $\mathbf{n}$ for any randomly arriving customer conditional on $v_i > v_{-i}$. Using Bayes' theorem, we have

$$\Pr[v_1 > v_2 \mid \mathbf{n}, s; \sigma] = \begin{cases} \dfrac{g\pi_s(\mathbf{n},\sigma)}{g\pi_s(\mathbf{n},\sigma)+(1-g)\pi_{-s}(\mathbf{n},\sigma)} & v_1 > v_2, \\[2ex] \dfrac{(1-g)\pi_{-s}(\mathbf{n},\sigma)}{g\pi_s(\mathbf{n},\sigma)+(1-g)\pi_{-s}(\mathbf{n},\sigma)} & \text{otherwise.} \end{cases} \tag{1}$$

After observing $(\mathbf{n}, s)$, the customer updates her prior expected service value for both service providers:

$$E[v_r \mid \mathbf{n}, s; \sigma] = \frac{g\pi_s(\mathbf{n},\sigma)E[v_r \mid v_1 > v_2] + (1-g)\pi_{-s}(\mathbf{n})E[v_r \mid v_2 > v_1]}{g\pi_s(\mathbf{n},\sigma)+(1-g)\pi_{-s}(\mathbf{n},\sigma)}, \quad r,\ s \in \{1, 2\},$$

where $E[v_1 \mid v_1 > v_2] = E[v_2 \mid v_1 < v_2] = v_h$ and $E[v_1 \mid v_1 < v_2] = E[v_2 \mid v_2 < v_1] = v_l$. Notice that only v_h and v_l are relevant for a risk-neutral customer. Consequently, any symmetric distribution $F(v_1, v_2)$ over a discrete or continuous domain with the same conditional expectations will lead to the same equilibrium. Note that we have assumed that $v_l > 0$, which, along with the absence of waiting costs, allows us to exclude balking action from $\mathcal{A}$.

3.4. Customer Equilibrium

Let σ_i be customer i's strategy. Fix the strategy of all customers $j \neq i$ at σ. From (1), customer i's belief of the service value upon observing $(s, \mathbf{n})$ is $\Pr[v_1 > v_2 \mid \mathbf{n}, s, \sigma]$. Let $\mathrm{BR}(\sigma)$ be the best response of a customer to σ. Then, $\sigma_i \in \mathrm{BR}(\sigma)$ if and only if

$$\begin{cases} E[v_s \mid \mathbf{n}, s, \sigma] \geq E[v_{-s} \mid \mathbf{n}, s, \sigma] \;\Rightarrow\; \sigma_i(s, s, \mathbf{n}) = 1, \\ E[v_{-s} \mid \mathbf{n}, s, \sigma] \geq E[v_s \mid \mathbf{n}, s, \sigma] \;\Rightarrow\; \sigma_i(-s, s, \mathbf{n}) = 1. \end{cases} \tag{2}$$

σ^* is a pure-strategy symmetric Nash equilibrium if $\sigma^* \in \mathrm{BR}(\sigma^*)$. It is apparent from the definition of best responses that the magnitude of the difference in service valuations does not play a role in affecting customers' decisions and, consequently, the equilibrium joining behavior. This is especially true for customers without waiting costs. We indicate $a^*(s, \mathbf{n})$ as the equilibrium action in a pure-strategy equilibrium upon observing $(s, \mathbf{n})$. A mixed strategy is determined analogously. For convenience of notation, we drop the dependency of the long-run probabilities and beliefs on σ. In the next section, we characterize σ^*.

4. Model Analysis for Symmetric Service Rates

4.1. Road Map of the Analysis

In this section, we first derive conditions for the customers' queue-joining strategy to be an equilibrium strategy when the service providers have identical service rates. To achieve this result, we begin with a restricted set of (pure) *threshold* strategies that examine "fixed queue-length differences." We show that the following strategy is an equilibrium strategy: Customers follow their signals when the queue lengths are equal, and follow the longer queue otherwise. Relaxing the restricted set of strategies to include all possible threshold strategies, any equilibrium threshold strategy is identical to the aforementioned strategy at all recurrent states. We find that our equilibrium result is robust when extended to mixed equilibrium strategies, multiple servers, and the possibility that service providers may go out of business if they operate a long time without customers. We describe the challenges of a full waiting cost analysis in §5.3. Finally, we expand the analysis of symmetric service rates to include correlated asymmetric services in §6.

4.2. Equilibrium Customer Strategies

We begin by describing the actions of a valuation-maximizing rational customer. We derive the conditions for each action every rational customer would take at each state and each signal. A strategy is a customer action at each state $\mathbf{n}$ and each private signal s (i.e., join queue 1, Join a queue according to the signal, join queue 2). Based on the customer's action at each state, we can construct a two-dimensional birth and death process and the corresponding steady-state probabilities for all states of the system. A strategy is an equilibrium strategy if it gives rise to stationary probabilities that support the existence of the strategy at each state. To reduce the notational burden, we drop the dependency of $\pi_i(\mathbf{n}; \sigma)$ on σ. As defined before, $\pi_1(\mathbf{n})$ is the probability that the system is in state $\mathbf{n}$ for any randomly arriving customer conditional on $v_1 > v_2$, and $\pi_2(\mathbf{n})$ can be defined in a similar way. Let the likelihood ratio[1] $l(\mathbf{n}) = \pi_1(\mathbf{n})/\pi_2(\mathbf{n})$ be the likelihood that server 1 is better than server 2, given that the state of the system is $\mathbf{n}$. Then, we can show that

Lemma 1. *The equilibrium strategy satisfies*

$$a^*(1, \mathbf{n}) = 1 \ (2) \quad \textit{if } l(\mathbf{n}) \geq (\leq) \, \frac{1-g}{g};$$

$$a^*(2, \mathbf{n}) = 1 \ (2) \quad \textit{if } l(\mathbf{n}) \geq (\leq) \, \frac{g}{1-g}.$$

At equality, customers are indifferent between queues 1 *and* 2.

Lemma 1 relates the choice of the customer (whether to choose queue 1 or queue 2) to the observed signal, state of the system (defined by the

[1] Note $l(\mathbf{n}) \geq 0$ wherever it exists. When $\pi_1(\mathbf{n}) > 0$, $\pi_2(\mathbf{n}) = 0$, $l(\mathbf{n}) = \pi_1(\mathbf{n})/\pi_2(\mathbf{n}) = \infty$. When $\pi_1(\mathbf{n}) = 0$ and $\pi_2(\mathbf{n}) = 0$, $l(\mathbf{n})$ is indeterminate.

number of customers waiting in each queue), and the likelihood ratio at that state. A customer chooses to join queue 1 after seeing signal 1 if the likelihood ratio is greater than $(1-g)/g$. If the likelihood ratio is less than $(1-g)/g$, the customer chooses queue 2. The above equilibrium conditions combine the imperfect information gained from the signal about the qualities of the servers, and information about the service provider acquired through the choices that *other* customers have made.

Lemma 1 can be understood using joint probabilities. For example, a customer joins queue 1 upon observing signal 1 or $a(1, \mathbf{n}) = 1$, when $l(\mathbf{n}) \geq (1-g)/g$, i.e., when $g\pi_1(\mathbf{n}) \geq (1-g)\pi_2(\mathbf{n})$. In other words, the customer arriving at some state $\mathbf{n}$ on seeing signal 1 follows queue 1 rationally, when the probability of being at the "right" state of nature *and* seeing the "true" signal (1 is better than 2 and this signal is true) is higher than the probability of being in the "wrong" state (server 2 is better than server 1) and seeing a false signal (signal 1). Conversely, $a(1, \mathbf{n}) = 2$ when $l(\mathbf{n}) \leq (1-g)/g$, i.e., when $g\pi_1(\mathbf{n}) \leq (1-g)\pi_2(\mathbf{n})$.

Arriving at state $\mathbf{n}$, the customer has the following choices: follow the shorter queue; follow the longer queue independent of the signal realized; or, finally, just follow one's own private signal. Now we are ready to define herding in our model. A customer *herds* at $\mathbf{n}$ when she joins the *longer* queue at $\mathbf{n}$ independent of her signal.[2] The customer is herding at that state by ignoring her private signal and joining the longer queue based on the public information.

Now we can postulate the structure of different strategies and analyze the equilibrium behavior. Clearly, there are several possible rational strategies that exist in this game.

4.2.1. Fixed-Threshold Strategies. We initially focus on a class of strategies that are *fixed queue difference threshold* strategies. Consider the following strategy for a customer arriving at some state $\mathbf{n}$: The customer follows her own private signal when $|n_1 - n_2| \leq B$ and follows the longest queue at all other states. A consumer arrives at some state and observes the queue lengths at the servers. If the difference between the queue lengths is strictly greater than B, the customer joins the longer queue. Otherwise, she follows her signal. We denote such a strategy by $\mathcal{A}^B$. B denotes the threshold (queue difference) above which each arriving customer will join the longer queue. The term *fixed* refers to the fact that the queue-length difference thresholds are fixed at B regardless of the length of the shorter queue. In other words, regardless of the state at which the customers arrive, if the queue-length difference is in the set $\{-B, -(B-1), \ldots, 0, \ldots, B-1, B\}$, the customers follow their own private signal. Because the threshold always refers to the difference in queue lengths, we refer to fixed queue difference threshold strategies as just *fixed-threshold* strategies, whenever it is unambiguous.

Fixed-threshold strategies might be a reasonable behavioral assumption, because the assumption reduces the queue-selection problem of a customer to a simple comparison between a single decision parameter (namely, the queue-length difference) and her own private signal. In §4.2.3, we generalize the fixed-threshold strategies.

Note that when $B = \infty$, customers always follow their private signal and ignore the queue-length information. Notice that in a game in which customers would *not* observe the queue length, customers would follow their signal because this is the only differentiator of the two service providers. We now explore whether this strategy is an equilibrium strategy when queue lengths are observable:

Proposition 2. *It is never an equilibrium strategy for customers to always follow their private signal* (*i.e., to always choose a queue independent of the queue lengths*).

In other words, when service providers are symmetric in service rates, ignoring queue lengths ($\mathcal{A}^\infty$) is never an equilibrium strategy. This result holds independent of the strength of the private signal as long as the signal is imperfect or noisy ($g < 1$). If every customer follows his or her signal, we find that the longer queue forms in front of the higher-valuation provider because more customers get the right signal (because $g > 1/2$). Hence, an arriving customer has an incentive to deviate from just following her signal, and would join the longer queue to improve her expected

[2] Our definition of herding is slightly different from the classical herding literature (see, e.g., Chamley 2003, pp. 64–65). We consider the decision of an arriving customer in a steady state, whereas the classical literature considers a transient regime.

service valuation. Therefore, $\mathscr{A}^{\infty}$ fails to be an equilibrium strategy. Proposition 2 captures the argument that customer decisions are influenced by externalities. It is *never* an equilibrium strategy to completely ignore the available public information pertaining to the decisions that other customers have made. This emphasizes the importance of studying the value of information accrued from the available public information, such as queue lengths.

We now examine which fixed queue difference threshold strategies are in equilibrium in the following proposition. At any state of the system, if following one's private signal is an equilibrium strategy, then at that state the customer must have likelihood ratios between $(1-g)/g$ and $g/(1-g)$. Therefore, for a fixed-threshold strategy B to hold in equilibrium, the likelihood ratios in all the states within the queue difference B should be between $(1-g)/g$ and $g/(1-g)$. Within the fixed threshold, if these likelihood conditions do not hold, the customers have an incentive to not follow their signal. The likelihood ratios must be greater than $g/(1-g)$ if the customers ignore the signal and follow the longer queue outside the fixed queue difference.

Proposition 3. *It is never an equilibrium strategy for customers to join the longest queue only if the queue-length difference is greater than some positive threshold $B \geq 1$, and follow their private signal otherwise.*

In other words, Proposition 3 implies that $\mathscr{A}^{B}$ is never an equilibrium strategy for any $B \geq 1$. Strategies in which customers follow their private signal within the fixed queue difference (of size one or more) generate stationary probabilities (and likelihood ratios) that do not satisfy the condition required for following the signal (from Lemma 1) in all states whose queue difference is bounded by the fixed threshold. Hence, they fail to hold in equilibrium. For instance, suppose all the customers follow the strategy $\mathscr{A}^{1}$, i.e., each customer follows her private signal if the queue lengths are equal or differ by one. Otherwise, she chooses the longer queue. Solving the two-dimensional birth and death process corresponding to the actions taken in the strategy $\mathscr{A}^{1}$, we obtain the stationary probabilities of being at each state. Using these probabilities, we can obtain the likelihood ratios at each state. If $\mathscr{A}^{1}$ is an equilibrium strategy, the likelihood ratios at each state should in turn satisfy the requirements according to Lemma 1 at each state. However, they do not.

The intuition for why $\mathscr{A}^{1}$ is not an equilibrium strategy is the following: Consider a customer arriving at the system and observing the state $(1,0)$ and following her own private signal. If the customer sees signal 1, the customer follows the signal because his signal is consistent with the longer queue. If the customer receives signal 2 at $(1,0)$, she will follow her signal rationally only if the probability of being at the "right" state of nature *and* seeing the "false" signal (server 1 is better than server 2, but the signal says 2) is lower than the probability of being in the "wrong" state (server 2 is better than server 1) and seeing a "true" signal (signal 2) (i.e., $(1-g)\pi_1(1,0) \leq g\pi_2(1,0)$ or simply, $l(1,0) \leq g/(1-g)$).

Based on the customer actions at each state under the strategy $\mathscr{A}^{1}$, we derive the probabilities of observing the state $(1,0)$ when server 1 is better $(\pi_1(1,0))$ and when server 2 is better $(\pi_2(1,0))$. Based on these probabilities, we show in the proof of Proposition 3 that the likelihood ratio at $(1,0)$ violates the aforementioned condition (i.e., we show that $l(1,0) > g/(1-g)$). Hence, through Lemma 1 we find that the best response of a customer arriving at $(1,0)$ is to follow the longer queue at $(1,0)$ if everyone else follows the strategy $\mathscr{A}^{1}$. Each customer arriving at the state $(1,0)$ has an incentive to deviate from the strategy $\mathscr{A}^{1}$ and improve her payoff, given that every other customer follows the $\mathscr{A}^{1}$ strategy. Therefore, the $\mathscr{A}^{1}$ strategy fails to be an equilibrium strategy. Similar arguments can be made for all other fixed-threshold strategies by showing that at the state $(1,0)$ or some other state, an arriving customer has an incentive to deviate from $\mathscr{A}^{B}$. Finally, note that when all customers follow the fixed-threshold strategy with $B \geq 1$, the customers can queue up in front of both servers with nonzero probability.

We consider the sole remaining candidate for the equilibrium strategy in the class of fixed queue difference threshold strategies, $\mathscr{A}^{0}$, in Proposition 4.

Proposition 4. *It is an equilibrium strategy for customers to follow their signal when the queue lengths are equal, and join the longer queue otherwise.*

The result implies that $\mathscr{A}^{0}$ strategy is supported in equilibrium. Because $g > 1/2$ whenever the queue lengths are equal, the customers join the better server with higher probability (because all customers follow their own signal, and the signal is accurate with probability $g > 1/2$). Once a particular queue becomes longer than the other queue, it keeps growing with every new arrival until eventual service departures (note that $\mu > \lambda$) from the longer queue bring the queue lengths back to being equal. Therefore, on average, the better service provider has a longer queue. The notion that if all the arriving customers join the longer queue they have a higher likelihood of getting better service value is true. It can be shown that $l(n, 0)$ (or $l(0, n)$) is equal to $g/(1-g)$ (or $(1-g)/g$) $\forall n$ under $\mathscr{A}^0$. Hence, the customers weakly prefer the longer queue. This is because under $\mathscr{A}^0$, when the queue is not empty, the private information held by a customer is not reflected in her queue choice. In other words, the herd behavior *dilutes* the information contained in the queue. Only when a customer arrives at an empty system, is the private information she holds revealed (because she follows her signal at that state). This naturally raises a question as to whether the customers could randomize their decisions in equilibrium.

4.2.2. Mixed Strategies. Our next result asserts that the customers will never mix between the strategies of joining the longer queue and following their signal when they are indifferent between the valuations gained. Let p_k be the probability that the customer follows the signal at state $(k, 0)$ (he ignores it and joins the longer queue with probability $1-p_k$). Let the actions be symmetric in state $(0, k)$. We show such a mixing strategy will not be an equilibrium strategy.

Proposition 5. *Under $\mathscr{A}^0$, it is never an equilibrium strategy for customers to mix at all states in which they are indifferent. There exists at least one state at which customers will always follow the longer queue as a best response.*

The notion that the customers join the longer queue, ignoring shorter queue lengths, is appealing. We note that in many real-life occasions, such as when selecting restaurants, there is some evidence for customers choosing the longest queue (Becker 1991, Hill 2007).

4.2.3. General Threshold Strategies. It is certainly a restriction to explore only fixed-threshold strategies, although such a restriction allows us to reduce the complexity of the equilibrium analysis. Consider the following two states $(2, 0)$ and $(102, 100)$. Although queue-length differences in those states are the same, the queue-length difference may be of much more significance at the state $(2, 0)$ than at the state $(102, 100)$. Customers might join the longer queue at $(2, 0)$ and not at $(102, 100)$. We address this issue by considering *general threshold* strategies where the queue-joining behavior not only depends on the queue-length differences alone, but also on the length of the individual queues. Under the class of general threshold strategies, the customers join the longer queue when the length of the longer queue is greater than $T_k + k$, where k is the length of the shorter queue and T_k is a nonnegative integer (no other structure is imposed on T_k). Define the class of threshold strategy where each strategy is defined by $\mathscr{S}^{\{T_0, T_1, \ldots\}}$ where T: $\mathbb{N} \mapsto \mathbb{N}$ as a threshold strategy. Clearly, fixed queue difference strategy $\mathscr{A}^B$ is a threshold strategy where $T_k = B$ for all k. In other words, $\mathscr{A}^B \equiv \mathscr{S}^{\{B, B, \ldots\}}$. In the following proposition, we show that the equilibrium threshold strategies are almost surely identical to the $\mathscr{A}^0$ strategy.

Proposition 6. *Threshold strategies are in equilibrium only if $T_0 = 0$, i.e., when customers always join the longer queue when the shorter queue is empty.*

Hence, equilibrium threshold strategies are identical to the $\mathscr{A}^0$ strategy at all nonzero probability states. We observe that in equilibrium, regardless of the values of threshold levels that exist at states when the shorter queue $k \geq 1$, the best response for a customer arriving at $(1, 0)$ or $(0, 1)$ is to join the longer queue. Thus, just as with $\mathscr{A}^0$, when customers queue in front of one server, the other server is empty. Therefore, the two queues can never develop simultaneously in steady state.

4.3. Market Implications of the Equilibrium Strategies

We observe that among all the strategies where the customers follow the longer queue above a threshold, the unique best-response strategy at $(1, 0)$ and $(0, 1)$ is to follow the longer queue. The results we derived in the section imply the following corollary.

Corllary 7. *The better server (of the two servers) is busy (and the other server is empty) during a fraction ρg*

of time. The worse server is busy during a fraction $\rho(1-g)$ of time.

It follows from Corollary 7 that the market share for the high-quality service provider is equal to the strength of the signal. The low-quality service provider captures the residual market. Somewhat surprisingly, the service rate does not play a role for the market share. It only determines the total traffic ρ to both servers. This is because no customers ever balk at the system. The same market share (and busy times of each service provider) would be obtained if customers would completely ignore the queue-length information and just follow their private information.[3] However, the total number of customers in the system differs significantly: If customers completely ignored the queue-length information, the average number in the system would be much lower. This is intuitive: The herding reduces the effective system capacity because only one service provider is active at any point in time. The stationary equilibrium thus has the following properties:

1. At all times, one of the service providers is necessarily empty.

2. The system is empty with a nonzero probability. A customer that arrives when the queues are empty follows her private signal. Once she chooses a service, all future customers continue to choose the same service provider until the queue depletes to zero. These cycles then repeat with independent private signal realizations (for the customer arriving at the empty system), deciding which service provider's queue grows next.

3. Queues grow and decay in cycles. The proportion of time that a high-quality service provider is busy increases with the strength of the private signal. A service provider that is busy clears the queue at a rate that decreases in ρ. The long-run market share of the better firm, however, is equal to the signal strength g.

4.4. Survival Likelihood

A consequence of the analyzed equilibrium strategy is that all the customers line up at the service offered by one of the service providers. Consider the service provider that has an empty queue. We describe this service provider as "starving" for customers. So far we had assumed that starvation does not impact the service provider in our model. However, starving service providers do not earn revenues that cover their fixed costs. This may be especially important for start-ups (Archibald et al. 2002) that do not have easy access to capital markets. In this subsection, we model the possibility that a service provider goes out of business because of a lack of revenues from customer arrivals.

Let both service providers in the market be described by a 2-tuple (v, τ) where v is the valuation the service provider offers to the consumers from its service, $\tau \in [\underline{\tau}, \infty)$ is the survival parameter, and $\underline{\tau}$ is a lower bound on τ. τ denotes the time a starving service provider can survive without any revenues from arriving customers. τ is determined by the firm's ability to cover its fixed costs without earning revenues. After τ units of time starving for customers, the service provider goes out of business. Suppose (τ_1, τ_2) is distributed with a symmetric density function $\phi(\tau_1, \tau_2)$ in the market. In the following result, we show that $\mathscr{A}^0$ remains an equilibrium strategy, although starving service providers might go out of business.

Proposition 8. *Let a service provider i in the market be described by (v_i, τ_i), where v_i is the valuation offered to customers, τ_i denotes the time it can survive without any revenues from arriving customers, and $\tau_i \in [\underline{\tau}, \infty)$, $i \in \{1, 2\}$, $\underline{\tau} > 0$. Conditional on the presence of two servers in the market, strategy $\mathscr{A}^0$ is an equilibrium strategy. When there is one server in the market, all customers join the server.*

Proposition 8 extends the applicability of the results obtained earlier in this section. The intuition is that the observation by the customers that both service providers are in the market does not contain any additional information about the relative quality of the service providers. Once one service provider goes out of business, the customers in the market are forced to join the other one for the lack of an alternative. Hence, the queue-joining problem becomes trivial when there is only one firm in the market. Note that with probability one, the remaining service provider will eventually go out of business too, because it cannot be

[3] In that case, the arrival rate to the high (low)-quality service provider would be $g\lambda$ $((1-g)\lambda)$.

guaranteed that the idle period with a single service provider is less than τ.

5. Extensions of Modeling Assumptions

In the previous section we resorted to simplifying assumptions, primarily to obtain sharp analytical insights. In this section, we explore extensions of the base model. Specifically, we consider multiple servers, allow the priors to be asymmetric, and discuss the effect of waiting costs on the decision of a single customer in the system.

5.1. Multiserver Extension

We now model each service provider as a service station with N identical servers. In this extension, the customers arrive at restaurants or facilities that have multiple service providers (for example, think of a restaurant with N tables where each table hosts one customer and all the tables are served by waiting personnel with identical service speeds).

Proposition 9. *$\mathcal{A}^0$ is an equilibrium strategy when each service provider has N multiple servers.*

Proposition 9 notes that $\mathcal{A}^0$ continues to hold in equilibrium. We can apply our insights to two service providers offering services through several identical servers, or selling products with unknown valuation through identical retailing centers. It is also worth noting that customers continue to weakly prefer the longer queue when they are indifferent between valuations gained, just as in the single-server case.

5.2. Asymmetric Priors About Service Valuations

We assumed the initial priors of customers were symmetric, i.e., $\Pr[v_1 = v_h] = \Pr[v_1 = v_l] = 1/2$. However, customers might have asymmetric priors. Without loss of generality, let customers believe that service provider 1 is better with probability q_0; i.e., $\Pr[v_1 = v_h] = q_0$ and $\Pr[v_1 = v_l] = 1 - q_0$. We can rederive the likelihood ratio conditions similar to Lemma 1, now employing asymmetric priors, and show that the equilibrium strategy satisfies the following condition:

$$a^*(1, \mathbf{n}) = 1 \ (2) \quad \text{if } l(\mathbf{n}) \geq (\leq) \frac{1-g}{g}\frac{1-q_0}{q_0};$$

$$a^*(2, \mathbf{n}) = 1 \ (2) \quad \text{if } l(\mathbf{n}) \geq (\leq) \frac{g}{1-g}\frac{1-q_0}{q_0}.$$

At equality, customers are indifferent between queues 1 and 2. Exploring fixed-threshold strategies over the conditions, we find that

1. When $q_0 \geq g$, customers always join service provider 1 at all states.
2. When $q_0 \leq 1 - g$, customers always join service provider 2 at all states.
3. When $1 - g < q_0 < 1/2$ and $1/2 < q_0 < g$, none of the symmetric fixed-threshold strategies are in equilibrium. We conjecture that customers always join a longer queue identical to their initial prior at small queue-length differences, and join a longer queue opposite to their prior when those queues are much longer.

The first two results are not surprising. Suppose the strength of the private signal that the customers acquire is much weaker than their initial priors. Then the private signals that customers observe are not strong enough to overcome the initial priors that customers have against or for the service value from a service provider. Hence, all customers follow what the prior dictates. As a result, one service provider attracts all customers and the queue lengths become uninformative. When the priors are strictly asymmetric, but neither too strong nor too weak, we find that none of the symmetric fixed-threshold strategies are in equilibrium. In the case with weak initial priors, the equilibrium strategy will have characteristics of the case with symmetric priors and the case with extremely strong priors. In particular, we conjecture that if $q_0 > 1/2$, customers might follow longer queue 1 when the queue difference is small. They might follow queue 2 if it is the longer queue, but only if queue-length difference is sufficiently large.

5.3. Waiting Cost Challenges

The presence of waiting costs poses significant challenges to analysis. In this subsection, we discuss the underlying complexity of analyzing waiting costs, and defer a full analysis of the waiting cost problem to future research. The model with waiting cost introduces the customer to a trade-off she was not facing before: Observing a long queue indicates that more people chose this service, but it also indicates that the customer will have to wait longer for the service. This presence of additional trade-offs changes the properties of the equilibrium strategy we have derived.

An additional issue is that given some waiting cost, there is no tractable way to theoretically model a two-queue system with large buffer space and analyze the best response at each state. We illustrate this complexity by considering the best response of a single customer in the system.

Let c be the waiting cost per unit of time for customers arriving at a system containing two symmetric servers. Again, let $\mathscr{A} = \{0, 1, 2\}$ be the set of possible actions that the customer can take upon arrival; 1 represents joining server 1, 2 represents joining server 2, and 0 represents balking from the system.[4] A mixed strategy for this customer is then a mapping σ: $\mathscr{A} \times \mathscr{S} \times \mathbb{N}^2 \to [0, 1]$. Let $\sigma_j(a, s, \mathbf{n})$ be the probability that this customer j joins queue a after observing signal s and state $\mathbf{n}$ (with $\sum_{a\in\mathscr{A}} \sigma_j(a, s, \mathbf{n}) = 1$). Let σ_i be customer i's strategy. Fix the strategy of all customers $j \neq i$ at σ. Customer i's belief of the service value upon observing $(s, \mathbf{n})$ is $\Pr[v_1 > v_2 \mid \mathbf{n}, s, \sigma]$. The customer balks from the system if the expected valuation on seeing the signal is lower than the waiting cost. We can now characterize the best response of this customer (similar to Equation (2)). Let $\mathrm{BR}(\sigma)$ be the best response of a customer to σ. Then, $\sigma_i \in \mathrm{BR}(\sigma)$ if and only if

$$\begin{cases} \mathbb{E}[V_s \mid \mathbf{n}, s; \sigma] - c(n_s + 1)/\mu \\ \quad > \max(0, \mathbb{E}[V_{-s} \mid \mathbf{n}, s; \sigma] - c(n_{-s} + 1)/\mu) \\ \qquad \Rightarrow \quad \sigma_i(s, s, \mathbf{n}) = 1, \\ \mathbb{E}[V_{-s} \mid \mathbf{n}, s; \sigma] - c(n_{-s} + 1)/\mu \\ \quad > \max(0, \mathbb{E}[V_s \mid \mathbf{n}, s; \sigma] - c(n_s + 1)/\mu) \\ \qquad \Rightarrow \quad \sigma_i(-s, s, \mathbf{n}) = 1, \\ \max(\mathbb{E}[V_{-s} \mid \mathbf{n}, s; \sigma] - c(n_{-s} + 1)/\mu, \\ \qquad \mathbb{E}[V_s \mid \mathbf{n}, s; \sigma] - c(n_s + 1)/\mu) < 0 \\ \qquad\quad \Rightarrow \quad \sigma_i(0, s, \mathbf{n}) = 1. \end{cases} \tag{3}$$

There are two additional complications. Balking strategy is introduced in the action set, and the difference in expected valuations matters with respect to waiting costs. Given this information, customers might follow the longer queue, follow their signal, join the shorter queue, balk on seeing a contrary signal, or balk from the queues completely. However, given the steady-state probabilities, $\pi_1(\mathbf{n})$ and $\pi_2(\mathbf{n})$, their best responses can be immediately determined.

To the best of our knowledge, closed-form expressions for $\pi_1(\mathbf{n})$ and $\pi_2(\mathbf{n})$ for any strategy that could simplify the problem do not exist. For specifying the symmetric Nash equilibrium σ^*, we need $\sigma^* \in \mathrm{BR}(\sigma^*)$. Characterizing the equilibrium requires solving steady-state probabilities for a two-dimensional Markov chain for a given customer equilibrium strategy. This is challenging from a queuing point of view. Asymptotic approximations of the steady-state probabilities are known only for the special case where customers join shorter queues everywhere (Neuts 1981). This is only a subset of the strategies required in our equilibrium strategy determination problem. Even if one obtains the long-run probabilities for a given strategy, we still need to find the strategy that generates the long-run probabilities that are consistent with the strategy. An analysis of such a model is an intended future research direction. A finite waiting space model with waiting costs is considered in Veeraraghavan and Debo (2008).

6. Asymmetric and Unknown Service Rates

Often, certain services require careful attention to detail. For example, the assembly of a complex product and completion of some service might require a sequence of complex processes, skipping some of which will decrease the time of completion, but also decrease consumers' valuation of the finished product. For instance, customers might prefer to patronize a service or listen to advice that is carefully considered. This may be the case in industries where both the product and the production processes are new and evolving. In these cases, customers might have higher valuations for services that have taken more time (i.e., have a slower speed). On the other hand, for standardized or repetitive processes (such as photocopying, brewing coffee) that have a minimal number of operational steps, fast service could be

[4] Note that the action 0 does not refer to joining some real queue 0, but instead the action of balking from queues 1 and 2. We define balking as joining queue 0 for notational simplicity. Note that the customer does not observe n_0, the number of customers who have balked. This is commonly true: Customers often do not observe those customers who decided to balk from the queues.

an indication that all process steps were correct from the first time they were executed. When the process quality is less well controlled, some steps need to be redone, increasing the process lead time and generally degrading product quality. Longer processing time in such cases could be due to a failure, and would cause some reduction in valuation. Therefore, lower valuations are correlated with longer service times. In such cases, customers might have higher valuation for services and processes that have higher speed.

In this section, we explore the equilibrium strategies when service rates are unknown, but are either positively or negatively correlated to the service value. Define $\rho_i = \lambda/\mu_i$, $i = 1, 2$. The values μ_i are common knowledge; however, the customers do not know which service provider is the fastest and which one is the slowest. We do allow the customers to know whether the better server is faster or slower; however, the customers still do not know which is the better service provider. This assumption helps us to maintain symmetry and tractability. In the beginning of the game, the service values $v_1, v_2 \in \{v_h, v_l\}$ such that $v_1 \neq v_2$ are determined. The realization of (v_1, v_2) then immediately determines the service rates of both providers, depending on whether the better server is faster or the better server is slower. We consider the two possible cases: The case when the better server is slower (i.e., it has lower mean service rate) and the case when the better server is faster. Again, we will assume that server 1 is better, i.e., $v_1 > v_2$ without any loss of generality.

6.1. When the Better Service Provider Is Slower

In cases when careful attention is required to provide better service, the customers know that the better server is generally slower; i.e., they know that when $v_1 > v_2$, then $\rho_1 > \rho_2$, and similarly when $v_1 < v_2$, then $\rho_1 < \rho_2$. When the better service provider is slower, then we show the following result.

Proposition 10. *$\mathscr{A}^0$ is an equilibrium strategy under asymmetric service rates $\forall g \in (1/2, 1)$. This holds true as long as the better service provider is slower; i.e., $1 > \rho_1 > \rho_2$.*

Proposition 10 is especially significant because it connects customer herding behavior to relative rates of service completions and signal precision. Recall from the discussion following Proposition 4 that with symmetric and known service rates, the customers may be indifferent between the two queues when their private signal indicates the shorter queue; i.e., in the symmetric case, in equilibrium the customers only weakly prefer the longer queue. In contrast, here the customers *strictly* prefer joining the longer queue. The longer queue is long because either more customers joined the queue, or the service speed at the server is slower. In both cases, the customer infers that the service provider is likely to provide higher service valuation. The proposition further assists us in specifying the market shares of service providers.

Corllary 11. *Conditional on $v_1 > v_2$ (therefore $1 > \rho_1 > \rho_2$), server 1 is busy (and server 2 is empty) during a fraction $g\rho_1(1-\rho_2)/(1-(1-g)\rho_1 - g\rho_2)$ of time. Server 2 is busy (server 1 is empty) during a fraction $(1-g) \cdot \rho_2(1-\rho_1)/(1-(1-g)\rho_1 - g\rho_2)$ of time.*

Recall from the symmetric service rate case that the market share of the high-quality service provider was g. With the asymmetric service rates case, the market share of the high-quality server (server 1) is $g/((1-g) \cdot (\rho_2(1-\rho_1)/(\rho_1(1-\rho_2))) + g) > g$ (because $\rho_1 > \rho_2$). Note that the better server always takes a strictly higher market share than in the symmetric case. Furthermore, the market share is increasing in signal precision g and decreasing in its service rate μ_1. Our results imply that ex ante service rate uncertainty is beneficial for the high-quality server as long as better-quality service is associated with longer service time.

Recall from Corollary 7 that service rates when they are known and symmetric had no effect on the market shares of the servers. The higher-quality server captures a market share proportional to the signal strength. Corollary 11 generalizes this insight. Service rate uncertainty, provided that the better server is slower, actually increases the market share of the high-quality server. In other words, the uncertainty about service rates provides an alternative conduit for the high-quality server to have even higher market share without increasing the signal strength g. This is due to the positive correlation between high value and a long queue, and thus is particularly crucial in markets where the production process or service is known to be complex and time consuming.

6.2. When the Better Service Provider Is Faster

When waiting for low-skill standardized processes, customers know that the better service is generally faster. That is, when $v_1 > v_2$, then $\rho_1 < \rho_2$, and similarly $v_1 < v_2$ implies that $\rho_1 > \rho_2$. We obtain the following equilibrium strategy in the special case when $g = \rho_2/(\rho_1 + \rho_2)$.

Proposition 12. *If and only if $g = \rho_2/(\rho_1 + \rho_2)$, then $\mathcal{A}^\infty$ is an equilibrium strategy. The customers never herd at any state, i.e., customers always follow their signals regardless of queue lengths.*

Proposition 12 provides a (knife-edge) condition under which it is an equilibrium strategy for customers to ignore queue-length information completely. The proposition underlines the significance of the service process with respect to customer herding: Even though the queue lengths could contain potential information about the service value, we notice that uncertainty about the service rate, especially when the high-quality server provides fast service, destroys the service value information. In the case of Proposition 12, the signal strength is equal to the normalized speed of the better server. We examine a customer who arrives at, e.g., a state $(10, 5)$, and sees signal 2. The customer rationalizes from the public information that it is equally likely for server 1 to be better than server 2, and vice versa. This is because $l(10, 5) = 1$. She figures that because every customer has followed her signal, therefore at least 10 customers have received signal 1 and 5 customers have received signal 2 in the past. This might imply server 1 is better, but she knows that server 1 is also faster. Hence, it is also likely that more departures have occurred from server 1, and therefore it is likely to be less busy. Because every customer follows his or her signal, we have two independent M/M/1 queues, with load factor $g\rho_1$ at the faster server and $(1-g)\rho_2$ at the slower server. Using the steady-state probabilities for both queues, we find the likelihood ratio at some state $\mathbf{n}$ is $l(\mathbf{n}) = (g\rho_1/((1-g)\rho_2))^{n_2 - n_1}$. When the signal strength is such that $g = \rho_2/(\rho_1 + \rho_2)$, we have $l(\mathbf{n}) = (\rho_2/\rho_1 \cdot \rho_1/\rho_2)^{n_2-n_1} = 1\ \forall\, \mathbf{n}$. Because the likelihood ratios satisfy the equilibrium conditions (from Lemma 1) for following one's signal at all states (i.e., $(1-g)/g < l(\mathbf{n}) < g/(1-g)\ \forall\, \mathbf{n}$), the strategy $\mathcal{A}^\infty$ is an equilibrium strategy. In other words, at every state, both the longer queue and the shorter queue are equally likely to lead to the higher-quality service provider. No updating of her private belief occurs from the queue lengths, and she follows her signal. Note that this result is striking because it contrasts with the result of Proposition 2, in which the customers never follow their private signals when the service rates are known.

We provide limiting results when server 1 is much faster or when server 2 is much slower than arrival rates in the following proposition.

Proposition 13. 1. *If $1 \approx \rho_2 > \rho_1$, then none of the longer queue-joining finite fixed-threshold strategies $\mathcal{A}^k$ $\forall\, k \in N$ are in equilibrium.*

2. *If $\rho_1 \approx 0 \ll \rho_2$, then $\forall\, g \in (1/2, 1)$ in equilibrium customers join the shortest queue.*

Proposition 13.1 reveals that none of the finite fixed-threshold strategies are in equilibrium when the faster server is better. In general, the equilibrium strategy might be very complex. Customers might join shorter queues at some states and longer queues at some other states; thus, a nonthreshold strategy might be an equilibrium strategy for all customers. Proposition 13.2 indicates that shortest queue-joining behavior may occur despite the positive information externalities of joining the longest queue. This is because information from the service rate provides a much stronger signal than customers' private signal themselves. We thus provide a crucial result: The signal from faster service rates in the form of shorter queues could be a much stronger public signal than a customer's own private belief at some states. The customers ignore their private signal, and instead of joining longer queues, they might join shorter queues.

In summary, when the worse service provider is slower, longer-queue joining behavior may not always occur (unlike the case when the better server is slower). We provided conditions when customers might not herd at all. By examining limiting cases, we showed that customers might always join the shortest queue if the faster server is very fast. We also established that customers join shorter queues surely at some states, when the load factor approaches one. Notice that we have obtained significantly different results from BHW, who do not consider an operational context when studying herd behavior. Queues are a natural restriction of the observed histories. Service rate uncertainty can either reinforce or destroy herd behavior.

7. Conclusion and Future Direction

In our model, customers face a choice between two service providers with identical service rates and unknown service value. Customers are endowed with imperfect private information about which service provider is better. They also observe the queue lengths upon arrival. We have shown that customers may weakly prefer to completely ignore their private signal (when it does not indicate the longest queue), and join the longer queue even if the longer queue has only one additional customer.

Furthermore, if the better service provider has slower service rates, we show that customers strictly prefer to always join the longer queue. This equilibrium behavior has a pronounced effect on the dynamics of how a service provider's consumer base develops and decays. Existing research (e.g., Becker 1991) postulates that customers enjoy a show such as *Cats* more because they prefer it more when a lot of other customers also liked the show. We propose that a large number of customers might go and see a show like *Cats*, even if they do not surely know if it is good, because they do know that many other customers have seen it and therefore rationalize that it *must* be good. There exists a rational explanation from the perspective of each consumer that makes it appealing for him or her to follow others' actions, even if this means ignoring his or her private signal. We have shown that such queue-joining strategy is a consistent and rational strategy to employ while choosing between servers with uncertain quality.

What about fickleness of demand? Can there be episodes during which customers just do not patronize a service provider, even though neither the service provider nor its competitor has altered its operations? From our results, we know that congestion at a service provider exists as another service provider starves. Once the queue at the congested service provider trickles to zero, things might improve for the previously starving service provider, with positive probability. We show that such epochs exist in equilibrium, and starving cycles might alternate.

There is also the interesting role of fortuitous factors (due to signal realization when the queues are empty) in determining the temporary success (or failure) of a service provider. Success or failure of a service provider is a function of the realization of the private signal of the "first" taster. This observation also implies the underlying value of advertising and wooing "early adopters" or "taste makers" or investing in better signalling. Building on this fortuitous "first buyer" occurrence, the service provider can continue attracting customers through long queues even if the service provider does not provide service with the highest valuation in the market.

Only when queues are uninformative (e.g., because they are equally long), an arriving customer decides solely based on his private assessment (which may be incorrect). Such a choice in turn decides which service provider will become successful in the short run. Path dependencies therefore play a significant role in determining how customers choose between services, and also affect how successful a service provider is in the market.

One of our results, the cycles of high demand followed by low demand, may also provide some insights into differences between market dynamics at large movie theater multiplexes and theaters/Broadway shows. The length of the cycles of high demand is determined by the seat capacity. For theater shows, the capacity of the theater is usually limited. Hence, the successful runs are long (unfortunately, runs without demand also last long, and a show may not survive). As multiplexes typically have large capacity, the cycles are much shorter than observed in theaters. Therefore, more shifts in demand for movies should happen more frequently than for theaters.

Our analysis also suggests that the market share of the better service provider may increase when there is uncertainty about the service rate that is negatively correlated with the service value. There is possibly an incentive for servers to mask the service process in order to increase the uncertainty and make queues doubly informative. Furthermore, a better service provider gains market share from improving customer information (strength of the signal) through promotion and advertising and from reducing its service rate.

7.1. Applying Our Results

Our service selection model is based on three key observations: (1) Arrival and service completion are stochastic and they create queue dynamics. (2) The queue lengths upon arrival are a natural truncation of the system history for arriving customers. (3) Customers infer information about the service quality from observed congestion levels (queue lengths) and act accordingly. These observations state in essence that customers do not make a service selection decision in a "vacuum," but they are influenced by the "level of business activity" they observe. The equilibrium analysis of our model in this context leads to property that while one firm attracts customers, the other firm is idling. It leads immediately to the questions: Do we observe such behavior in the real world? And, if not, what are the other drivers of service selection in the real world? In real life, we conjecture that other confounding factors may play a role. We already analyzed the impact of uncertainty about the service rates. Such uncertainty can either reinforce herd behavior (when high service value is positively correlated with slow service rate), or annihilate it (in the opposite case). We also discussed the role that waiting costs may play. These factors significantly complicate the customer's value assessment. Furthermore, heterogeneity in terms of the preference for each of the service choices, in terms of signal strength, in terms of prior belief about the service providers, etc., significantly impact the information that can be derived from the congestion levels. Asymmetric priors about the service value and service rates may also change the equilibrium outcome. Finally, customers may not be Bayesian decision makers; they may follow some simple heuristics instead of Bayes' rule.

Based on our research, we find that congestion externalities in the presence of imperfect service quality information introduce

- Cycles of high demand followed by low demand,
- Negative correlation between demand at different service providers, and
- Fickleness: Which service provider is thriving more is determined by customers following their private information when both service systems are equally congested.

We conjecture that as long as the confounding factors are not too pronounced, the described phenomena should be observed in real business environments. Empirical verification of the described phenomena is important, but also challenging. It has been a consideration of many experimental economists recently. Manski (2000) discusses many identification problems that need to be overcome for a rigorous empirical verification of the interaction effects similar to those noted in Becker's model. Many of the results require careful collection and calibration of primary and secondary data. Such sophisticated data-modeling issues are currently beyond the scope of this paper. Hence, we think that our framework may be a rich one for further research.

7.2. Future Research

Because we examine the market in stationarity based on fixed service rates, our model applies better to scenarios like theaters, shows, etc., where customers are served at static service rates. Although shows can be made to run longer by deliberately underselling tickets, this idling of capacity will eventually be inferred by the population. Furthermore, the presence of empty seats in a theater might dilute the valuation of the product. Hence, we can envision scenarios in which service providers do not dynamically adjust the service. However, we recognize that such opportunities for idling services are widely existent, and note that this is a rich avenue for further exploration.

Most of our analysis was tractable because we could exploit the symmetry of service rates and prior beliefs in deriving the equilibrium strategies. Relaxing this symmetry poses analytical challenges because the interior states are likely to be visited in equilibrium. Although our model would continue to hold theoretically, for a complete analysis one has to resort to numerical approaches to characterize the equilibrium behavior in the asymmetric cases.

Finally, one significant challenge that remains is the analysis of waiting costs. We have assumed in this paper that there is no cost of waiting. Making this assumption helped us focus our attention fully on the *information value* contained in the length of a queue. Although we believe that customers might take others' decisions into account while choosing between queues, the analysis and the effect of waiting costs remain a pertinent issue. We hope that our insights in the case of herd behavior in queues without waiting costs can form a basis for further research and exploration of the issue posed by the cost of waiting.

Acknowledgments

This paper was previously titled "To Join the Shortest Queue or the Longest Queue: Inferring Service Quality from Queue Lengths." The authors sincerely thank Krishnan Anand, Gérard Cachon, Noah Gans, Serguei Netessine, Alan Scheller-Wolf, Anita Tucker, an anonymous associate editor, three reviewers, and the seminar participants at Carnegie Mellon University, New York University, Washington University, and the Wharton School, whose comments significantly improved this paper. An earlier version of the paper was a finalist on the INFORMS Junior Faculty Interest Group paper competition.

Appendix

Proof of Lemma 1. Before we prove Lemma 1, we obtain the expressions for updated valuations. First, let us consider the different action strategies that a customer can adopt. Let v_h and v_l be the updated valuations of the services based on the signal; i.e., $\mathrm{E}(v_i \mid v_i > v_{-i}) = v_h$ and $\mathrm{E}(v_{-i} \mid v_i > v_{-i}) = v_l$. The conditions we need to establish follow directly from the expected valuation expressions. The updated expected service value as a function of the signal s and queue lengths $\mathbf{n}$ is given by

$$E[v_s \mid \mathbf{n}, s] = \frac{g\pi_s(\mathbf{n})v_h + (1-g)\pi_{-s}(\mathbf{n})v_l}{g\pi_s(\mathbf{n}) + (1-g)\pi_{-s}(\mathbf{n})}, \quad \text{and}$$

$$E[v_{-s} \mid \mathbf{n}, s] = \frac{g\pi_s(\mathbf{n})v_l + (1-g)\pi_{-s}(\mathbf{n})v_h}{g\pi_s(\mathbf{n}) + (1-g)\pi_{-s}(\mathbf{n})}.$$

First, for notational convenience we suppress (n_1, n_2) in $\pi_i(n_1, n_2)$ whenever necessary and denote $\pi_i(n_1, n_2)$ as π_i for $i = 1, 2$. Similarly, $l(n_1, n_2) = \pi_1(n_1, n_2)/\pi_2(n_1, n_2)$ is represented as l when there is no ambiguity that a general state n_1, n_2 is being discussed. Expanding the notation for the valuations we have

$$E[v_1 \mid \mathbf{n}, 1] = \frac{l(g/(1-g))v_h + v_l}{(g/(1-g))l + 1} > 0,$$

$$E[v_2 \mid \mathbf{n}, 1] = \frac{l(g/(1-g))v_l + v_h}{(g/(1-g))l + 1} > 0,$$

$$E[v_1 \mid \mathbf{n}, 2] = \frac{(g/(1-g))v_l + lv_h}{g/(1-g) + l} > 0,$$

$$E[v_2 \mid \mathbf{n}, 2] = \frac{(g/(1-g))v_h + lv_l}{g/(1-g) + l} > 0.$$

The above equations provide the valuations from each queue based on the state of the system at arrival and the signal observed by the arriving customer.

Because $\underline{v} > 0$, the expected valuations of service from both the servers are nonnegative. Therefore, the customer never balks. We show that contradicting one's signal never occurs. Let us consider a customer who joins the queue by contradicting one's own signal. For this to occur, the customer should perceive queue 1 (2) to provide better service when she sees a signal value of 2 (1), i.e., $E[v_1 \mid \mathbf{n}, 1] \le E[v_2 \mid \mathbf{n}, 1]$ and $E[v_1 \mid \mathbf{n}, 2] \ge E[v_2 \mid \mathbf{n}, 2]$ should hold simultaneously.

$$E[v_1 \mid \mathbf{n}, 1] \le E[v_2 \mid \mathbf{n}, 1] \quad \text{and} \quad E[v_1 \mid \mathbf{n}, 2] \ge E[v_2 \mid \mathbf{n}, 2],$$

$$l\left(\frac{g}{1-g}\right)(v_h - v_l) \le (v_h - v_l) \quad \text{and}$$

$$l(v_h - v_l) \ge \left(\frac{g}{1-g}\right)(v_h - v_l),$$

$$l \le \left(\frac{1-g}{g}\right) \quad \text{and} \quad l \ge \left(\frac{g}{1-g}\right).$$

The conditions for contradicting signals are violated because $g > 1/2$, $(g/(1-g)) > ((1-g)/g)$. Hence, we can focus on actions that do not involve balking from the queues or contradicting one's signal.

Now we are ready to summarize the conditions for choosing one queue over another given a particular signal in terms of the likelihood ratio.

Signal 1: Customer would join queue 1; i.e., $a(1, \mathbf{n}) = 1$ if $E[(v_1 \mid \mathbf{n}, 1)] \ge 0$ and $E[v_1 \mid \mathbf{n}, 1] - E[v_2 \mid \mathbf{n}, 1] \ge 0$. We have $E[v_1 \mid \mathbf{n}, 1] > 0$ (which is clearly true). Furthermore, we need $E[v_1 \mid \mathbf{n}, 1] \ge E[v_2 \mid \mathbf{n}, 1]$,

$$l\left(\frac{g}{1-g}\right)v_h + v_l \ge l\left(\frac{g}{1-g}\right)v_l + v_h$$

$$\Rightarrow \quad l(\mathbf{n}) \ge \frac{1-g}{g} \quad \text{for } a(1, \mathbf{n}) = 1.$$

Clearly, therefore, $a^*(1, \mathbf{n}) = 2$ if $l(\mathbf{n}) \le (1-g)/g$. Thus, $a^*(1, \mathbf{n}) = 1$ (2) if $l(\mathbf{n}) \ge (\le)(1-g)/g$.

Signal 2: Similarly, $a^*(2, \mathbf{n}) = 1$ if $E[v_1 \mid \mathbf{n}, 2] > 0$ and $E[v_1 \mid \mathbf{n}, 1] \ge E[v_2 \mid \mathbf{n}, 1]$.

$$\left(\frac{g}{1-g}\right)v_l + lv_h \ge \left(\frac{g}{1-g}\right)v_h + lv_l$$

$$\Rightarrow \quad l(\mathbf{n}) \ge \frac{g}{1-g} \quad \text{for } a(2, \mathbf{n}) = 1.$$

Reversing the inequality we have $a^*(2, \mathbf{n}) = 2$ if $l(\mathbf{n}) \le g/(1-g)$. The customers are indifferent at equality. Thus, $a^*(2, \mathbf{n}) = 1$ (2) if $l(\mathbf{n}) \ge (\le) g/(1-g)$. □

Proof of Proposition 2. If server 1 provides service with higher valuation, a fraction g of the customers get the true signal and the rest of the customers get a false signal (that server 2 is better). If all customers choose to always follow their private signal, the resulting queue at server 1 will be an $M/M/1$ queue with mean arrival rate $g\lambda$. Similarly, the queue at server 2 would be an $M/M/1$ queue with arrival rate $(1-g)\lambda$. Furthermore, because the signals are iid, the queues are independent. From the equilibrium probability distribution of the $M/M/1$ queue, we obtain that $l(n_1, n_2) = (g/(1-g))^{n_1 - n_2}$. From Lemma 1, if this strategy is in equilibrium, we require $(g/(1-g))^{n_1 - n_2} \le (g/(1-g))$ $\forall n_1, n_2$, which is not true. Hence, it can never be an equilibrium strategy for a customer to always follow her private signal. □

Proof of Proposition 3. Without loss of generality, let server 1 be better than server 2 (i.e., $v_1 > v_2$) for the following analysis. Given a strategy (set of actions at every state by all customers), we write the steady-state transition equations and then solve for stationary probabilities. When server 2 provides higher valuation than server 1, we can write the same steady state by suitably replacing g with $1-g$.

Strategy $\mathscr{A}^b$ is the strategy in which all customers follow their private signal when $|n_1 - n_2| \le b$, i.e., when the queue lengths differ no more than b. The customers follow the longer queue if the lengths differ more than b. b might denote the degree of importance of the private signal versus the herding/longer-queue following behavior. If b is large, then in a large number of states, the customers value and follow their private signals instead of using the public information (i.e., observed crowd at the servers). Similarly, if b is small, the customers are more likely to follow the longer queue than to follow their signal in a large number of states.

Consider a strategy of fixed threshold b. We can write the steady-state balance equation at each state. Initially, let us consider the balance equations at the outermost states, i.e., all the states where the queue at server 2 is zero. The customers follow the longer queue when arriving at states $(k, 0)$ where $k \ge b+1$ and follow their own signal at all states $(k, 0)$ where $0 \le k \le b$. Recall that customers arrive at rate λ to the system and each server serves at rate μ. Also, let $\rho = \lambda/\mu$. Let $\pi_{m,n}$ be the long-run stationary probabilities of state (m, n) under some strategy when $v_1 > v_2$.

$$\pi_{k,0}(\lambda+\mu) = \pi_{k-1,0}\lambda + \pi_{k+1,0}\mu + \pi_{k,1}\mu \quad \forall k \ge b+2,$$

$$\pi_{k,0}(\lambda+\mu) = \pi_{k-1,0}\lambda g + \pi_{k+1,0}\mu + \pi_{k,1}\mu \quad \forall 1 \le k \le b+1.$$

Adding all the equations for $k = 1, \ldots, \infty$,

$$\begin{aligned}(\lambda+\mu)\sum_{k=1}^{\infty}\pi_{k,0} &\\ = \lambda g \sum_{k=1}^{b+1}\pi_{k-1,0} + \lambda \sum_{k=b+2}^{\infty}\pi_{k-1,0} + \mu\sum_{k=1}^{\infty}\pi_{k+1,0} + \mu\sum_{k=1}^{\infty}\pi_{k,1},\end{aligned} \tag{4}$$

$$\begin{aligned}(\lambda+\mu)\sum_{k=1}^{\infty}\pi_{k,0} &\\ = \lambda g \sum_{k=0}^{b}\pi_{k,0} + \lambda \sum_{k=b+1}^{\infty}\pi_{k,0} + \mu\sum_{k=2}^{\infty}\pi_{k,0} + \mu\sum_{k=1}^{\infty}\pi_{k,1},\end{aligned} \tag{5}$$

$$\lambda\sum_{k=1}^{b}\pi_{k,0} + \mu\pi_{1,0} = \lambda g\sum_{k=0}^{b}\pi_{k,0} + \mu\sum_{k=1}^{\infty}\pi_{k,1}, \tag{6}$$

$$\lambda(1-g)\sum_{k=1}^{b}\pi_{k,0} + \mu\pi_{1,0} = \lambda g\pi_{0,0} + \mu\sum_{k=1}^{\infty}\pi_{k,1}. \tag{7}$$

Suppose the strategy was A^1. When the fixed threshold is equal to one, Equation (7) becomes

$$(\lambda(1-g)+\mu)\pi_{1,0} = \lambda g\pi_{0,0} + \mu\sum_{k=1}^{\infty}\pi_{k,1}; \tag{8}$$

$$(\rho(1-g)+1)\pi_{1,0} = \rho g\pi_{0,0} + \sum_{k=1}^{\infty}\pi_{k,1}. \tag{9}$$

Writing a similar expression for $\pi_{0,1}$ we have

$$(\rho g+1)\pi_{0,1} = \rho(1-g)\pi_{0,0} + \sum_{k=1}^{\infty}\pi_{1,k}. \tag{10}$$

From the steady-state balance equations, we note that the probability transition matrix when $v_1 > v_2$ is a transpose of the transition matrix when $v_2 > v_1$. We have the likelihood ratio at $(1, 0)$:

$$\begin{aligned} l(1,0) = \frac{\pi_{1,0}(v_1>v_2)}{\pi_{1,0}(v_2>v_1)} &= \frac{\pi_{1,0}(v_1>v_2)}{\pi_{0,1}(v_1>v_2)} \\ &= \frac{\lambda g\pi_{0,0} + \mu\sum_{k=1}^{\infty}\pi_{k,1}}{\lambda(1-g)\pi_{0,0} + \mu\sum_{k=1}^{\infty}\pi_{1,k}} \frac{\lambda g+\mu}{\lambda(1-g)+\mu} \\ &= \frac{\lambda g\pi_{0,0} + \mu A}{\lambda(1-g)\pi_{0,0} + \mu B} \frac{\lambda g+\mu}{\lambda(1-g)+\mu}, \end{aligned}$$

where $A = \sum_{k=1}^{\infty}\pi_{k,1}$ and $B = \sum_{k=1}^{\infty}\pi_{1,k}$. We will show that $A > B$.

If strategy $\mathscr{A}^1$ holds, we require $l(k,1) > g/(1-g)$ $\forall k > 1$. Then we have

$$\pi_{k,1} \ge \frac{g}{1-g}\pi_{1,k},$$

which gives

$$\sum_{k\ge 2}\pi_{k,1} \ge \frac{g}{1-g}\sum_{k\ge 2}\pi_{1,k} > \sum_{k\ge 2}\pi_{1,k},$$

$$\sum_{k=1}^{\infty}\pi_{k,1} > \sum_{k=1}^{\infty}\pi_{1,k} \quad \text{i.e.} \Rightarrow \quad A > B.$$

Now consider the steady-state balance equation for the state $\pi_{1,0}$:

$$\mu(\pi_{1,0} + \pi_{0,1}) = \lambda\pi_{0,0}.$$

Substituting for $\pi_{1,0}$ and $\pi_{1,0}$ from Equations (8) and (10) in the above equation gives

$$\left[\frac{\rho g\pi_{0,0} + A}{\rho(1-g)+1}\right] + \left[\frac{\rho(1-g)\pi_{0,0} + B}{\rho g+1}\right] = \rho\pi_{0,0}$$

$$\Rightarrow \quad (\rho g+1)A + (\rho(1-g)+1)B = \rho^2 g(1-g)(\rho+2)\pi_{0,0}.$$

This gives $A > \rho^2 g(1-g)\pi_{0,0} > B$.

$$\begin{aligned} \frac{\pi_{1,0}}{\pi_{0,1}} &= \frac{\lambda g\pi_{0,0} + \mu\sum_{k=1}^{\infty}\pi_{k,1}}{\lambda(1-g)\pi_{0,0} + \mu\sum_{k=1}^{\infty}\pi_{1,k}} \left(\frac{\lambda g+\mu}{\lambda(1-g)+\mu}\right) \\ &= \frac{\rho g\pi_{0,0} + A}{\rho(1-g)\pi_{0,0} + B}\left(\frac{\rho g+1}{\rho(1-g)+1}\right) \\ &> \frac{\rho g\pi_{0,0} + \rho^2 g(1-g)\pi_{0,0}}{\rho(1-g)\pi_{0,0} + \rho^2 g(1-g)\pi_{0,0}}\left(\frac{\rho g+1}{\rho(1-g)+1}\right) \\ &= \frac{\rho g(1+\rho(1-g))\pi_{0,0}}{\rho(1-g)(1+\rho g)\pi_{0,0}}\left(\frac{1+\rho g}{1+\rho(1-g)}\right) = \frac{g}{1-g}. \end{aligned}$$

The condition for following the signal condition is violated at $(1, 0)$ (and at the state $(0, 1)$) if all customers follow A^1.

Therefore, A^1 is not an equilibrium strategy. The analysis for $\mathscr{A}^b$ is similar when thresholds $b > 1$, and therefore is deferred to the technical appendix. This concludes the proof. □

Proof of Proposition 4. First, we calculate the long-run equilibrium probabilities under always following the longer-queue strategy. Under this strategy, the customer follows the signal at (n, n) $\forall n$. At other states, she follows the longer queue. It can be seen that only states $(n_1, 0)$ and $(0, n_2)$ with $n_1, n_2 \geq 0$ will be recurrent states. $\pi_1(n_1, 0) = g\rho^{n_1}(1-\rho)$, $\pi_1(0, n_2) = (1-g)\rho^{n_2}(1-\rho)$, and $\pi_1(0, 0) = 1 - \rho$ satisfy the steady-state probability conditions. We can obtain a similar expression for π_2 and obtain

$$l(n_1, n_2) = \begin{cases} \dfrac{g}{1-g} & n_1 \geq 1,\ n_2 = 0, \\ 1 & n_1 = 0,\ n_2 = 0, \\ \dfrac{1-g}{g} & n_1 = 0,\ n_2 \geq 1. \end{cases}$$

Following the signal at $n_1 = 0$, $n_2 = 0$ is an equilibrium action if $(1-g)/g \leq l(n_1, n_2) \leq g/(1-g)$, which is satisfied because $(1-g)/g < 1 < g/(1-g)$. At all reachable states $(n_1, 0)$ $n_1 > 0$ we have $l(n_1, 0) = g/(1-g)$, where customers are indifferent and following the longer queue is consistent with the strategy. Similarly, the condition of following the longest queue when $n_2 \geq 1, n_1 = 0$ is also weakly satisfied.

Because $b = 0$, all the interior states are transient and the only recurrent states are $(n_1, 0)$ and $(0, n_2)$ with $n_1, n_2 \geq 0$. Hence, it follows that $\mathscr{A}^0$ is the only fixed-threshold equilibrium strategy. □

Proof of Proposition 5. Without loss of generality, let server 1 be better than server 2 (i.e., $v_1 > v_2$) again. Consider the class of strategies where customers mix between following the signal and following the longer queue, along all the recurrent states in $\mathscr{A}^0$ strategy. Let us stipulate that at state $(k, 0)$ (and $(0, k)$) the customers follow their signal with probability p_k $(0 \leq p_k \leq 1)$ or ignore their signal, and follow the longer queue with probability $1 - p_k$. Given a strategy (set of actions at every state by all customers), we write the steady-state transition equations and then solve for stationary probabilities. Using the derived probability distributions, regardless of the customer actions at the other (interior) states, we show that such a mixing cannot occur in equilibrium. Specifically, we show that there exist at least some k such that $\pi_{k,0}/\pi_{0,k} > g/(1-g)$. (For mixing to occur at that state, we require $\pi_{k,0}/\pi_{0,k} = g/(1-g)$.) Therefore, there is at least one state where customers deviate and join the longer queue. The details of how the inequality is derived are established in the online technical appendix. □

Proof of Proposition 6. The first part of the proof is similar to the proof of Proposition 3 in the paper, where higher fixed queue difference threshold strategies are ruled out. Consider any threshold strategy such that $T_0 \geq 1$. Note that when $T_0 \geq 1$, some interior states are recurrent. Writing the steady balance equations for states on the outer arm, we show in the technical appendix that there is some state *within* the threshold at the outer arm, where the best response of an arriving customer would be to NOT follow the signal if all other customers were to follow the signal within the threshold.

For the second part of the proof, note that when $T_0 = 0$, all the states (n_1, n_2) where $n_1 > 0$ and $n_2 > 0$ are transient for $\lambda < \mu$. Only recurrent states are $(n_1, 0)$ and $(0, n_2)$ for any $n_1 \geq 0$ or $n_2 \geq 0$. When $T_0 = 0$, the actions at all recurrent states are identical to the strategy $\mathscr{A}^0$. This completes the proof. □

Proof of Corollary 7. Follows directly from the proof of Proposition 4. □

Proof of Proposition 8. $\mathscr{A}^0$ is in equilibrium when firms survive only for a finite time without customers. The proof is based on conditioning the $\mathscr{A}^0$ strategy on the number of firms in the market. See the online technical appendix for the detailed derivation. □

Proof of Proposition 9. The proof for multiserver firms is similar to the argument made in the proof of Proposition 4 for single-server firms, and therefore, is deferred to the online technical appendix. Thus, we show that $\mathscr{A}^0$ continues to exist as an equilibrium strategy even when the providers have N multiple servers each. □

Proof of Proposition 10. Without loss of generality, let $v_1 > v_2$ so that $(\rho_1 = \rho_S) > (\rho_2 = \rho_F) > 0$. We aim to prove that $\mathscr{A}^0$ is an equilibrium strategy. Once again, we calculate the long-run equilibrium probabilities under always following the longer-queue strategy. The customer follows the signal at (n, n) $\forall n$. At other states, she follows the longer queue. It is evident that states $(n_1, 0)$ and $(0, n_2)$ with $n_1, n_2 \geq 0$ will be recurrent states. Calculating the steady-state probabilities, we get

$$\pi_1(0,0) = \frac{(1-\rho_1)(1-\rho_2)}{[1-(1-g)\rho_1 - g\rho_2]} = \frac{(1-\rho_S)(1-\rho_F)}{[1-(1-g)\rho_S - g\rho_F]},$$
$$\pi_1(n_1, 0) = g\rho_1^k \pi_1(0,0) \quad \forall n_1 \geq 1,$$
$$\pi_1(0, n_2) = (1-g)\rho_2^k \pi_1(0,0) \quad \forall n_2 \geq 1.$$

Similarly, when $v_2 > v_1$, we get

$$\pi_2(0,0) = \frac{(1-\rho_1)(1-\rho_2)}{[1-g\rho_1 - (1-g)\rho_2]} = \frac{(1-\rho_S)(1-\rho_F)}{[1-(1-g)\rho_S - g\rho_F]},$$
$$\pi_2(n_1, 0) = (1-g)\rho_2^k \pi_2(0,0) \quad \forall n_1 \geq 1,$$
$$\pi_2(0, n_2) = g\rho_1^k \pi_2(0,0) \quad \forall n_2 \geq 1.$$

Now consider likelihood ratios at all recurrent states;

$$l_{00} = \frac{\pi_1(0,0)}{\pi_2(0,0)} = \frac{[1-g\rho_1-(1-g)\rho_2]}{[1-(1-g)\rho_1-g\rho_2]}$$
$$= 1 \quad \text{follow signal,}$$

$$\begin{aligned} l_{k0} &= \frac{\pi_1(k,0)}{\pi_2(k,0)} = \frac{g\rho_S^k\pi_1(0,0)}{(1-g)\rho_F^k\pi_2(0,0)} = \left(\frac{g}{1-g}\right)\frac{\rho_S^k}{\rho_F^k} \\ &> \frac{g}{1-g} \quad \text{follow the longer queue for all } k, \\ l_{0k} &= \frac{\pi_1(0,k)}{\pi_2(0,k)} = \frac{(1-g)\rho_F^k\pi_1(0,0)}{g\rho_S^k\pi_2(0,0)} = \left(\frac{1-g}{g}\right)\frac{\rho_F^k}{\rho_S^k} \\ &< \frac{1-g}{g} \quad \text{follow the longer queue for all k.} \end{aligned}$$

Therefore, $\mathcal{A}^0$ is an equilibrium strategy and customers strictly prefer joining the longer queues. □

Proof of Corollary 11. Directly follows from the proof of Proposition 10. □

Proof of Proposition 12. Without loss of generality, let $v_1 > v_2$ so that $\rho_1 = \rho_F = (g/(1-g))\rho_2 = (g/(1-g))\rho_S$. We aim to prove that $\mathcal{A}^\infty$ is an equilibrium strategy where every customer follows his signal at all states. Under this strategy, we have two independent $M/M/1$ queues: one with arrival rate $g\lambda$ and service rate μ_1, and another queue with arrival rate $(1-g)\lambda$ and service rate μ_2. Calculating the steady-state probabilities when $v_1 > v_2$, and when $v_2 > v_1$, we get

$$\begin{aligned} l_{00} &= \frac{\pi_1(0,0)}{\pi_2(0,0)} = \frac{(g\rho_1)^0(1-g\rho_1)(1-(1-g)\rho_1)}{((1-g)\rho_1)^0(1-(1-g)\rho_1)} \\ &= \frac{(1-g\rho_F)(1-(1-g)\rho_S)}{(1-(1-g)\rho_S)(1-(1-g)\rho_F)} = 1, \\ l_{mn} &= \frac{\pi_1(m,n)}{\pi_2(m,n)} \\ &= \frac{(g\rho_F)^m(1-g\rho_F)}{((1-g)\rho_S)^m(1-(1-g)\rho_S)}\,\frac{((1-g)\rho_S)^n(1-(1-g)\rho_S)}{(g\rho_F)^n(1-g\rho_F)} \\ &= \left(\frac{g\rho_F}{(1-g)\rho_S}\right)^{m-n} = 1 \quad \forall\, m,\, n. \end{aligned}$$

Hence, all customers rationally follow the signal at every state. Therefore, $\mathcal{A}^\infty$ is an equilibrium strategy. Customers do not herd at any state of arrival, regardless of the difference between queue lengths. □

Proof of Proposition 13. For part (i), it is straightforward to show $\mathcal{A}^0$ is not an equilibrium strategy because $l_{k0} = \pi_1(k,0)/\pi_2(k,0) = g\rho_F^k\pi_1(0,0)/((1-g)\rho_S^k\pi_2(0,0)) = (g/(1-g))\rho_F^k/\rho_S^k < g/(1-g)$. We consider other fixed-threshold strategies $\mathcal{A}^b$ with thresholds $b \geq 0$. The detailed proof is provided in the technical appendix, but a sketch of the proof concept is provided here. We begin by summing up the probabilities of all the states along the ith diagonal. Because $\rho_2 \approx 1$, we have $\pi_{i,0} \to 0\ \forall\, i$. Because the service rates are high, the process behaves asymptotically as a birth and death process on both sides of the diagonal, with each state being one of "diagonals" (where the ith state is defined as the sum of the states $\bigcup_k (k+i, k)$). Then consider the asymptotic limiting expression for the bth upper and lower diagonals (the bth state in the birth and death process). We find that there is at least one state along the $(k+b+1, k)$ diagonal such that the customers' best response at that state is to follow their signal. (Strategy $\mathcal{A}^b$ requires that they follow the longer queue.) Therefore, $\mathcal{A}^b$ cannot be in equilibrium. Part (ii) is similarly proven by taking the load factor ρ asymptotically to zero. □

Electronic Companion

An electronic companion to this paper is available on the *Manufacturing & Service Operations Management* website (http://msom.pubs.informs.org/ecompanion.html).

References

Archibald, T. W., L. C. Thomas, J. M. Betts, R. B. Johnston. 2002. Should start-up companies be cautious? Inventory policies which maximise survival probabilities. *Management Sci.* **48**(9) 1161–1174.

Banerjee, A., D. Fudenberg. 2004. Word-of-mouth learning, games and economic behavior. *Games Econom. Behav.* **46**(1) 1–22.

Becker, G. 1991. A note on restaurant pricing and other examples of social influences on price. *J. Political Econom.* **99**(5) 1109–1116.

Bikhchandani, S., D. Hirshleifer, I. Welch. 1992. A theory of fads, fashion, custom and cultural change as information cascades. *J. Political Econom.* **100** 992–1026.

Callander, S., J. Horner. 2006. The wisdom of the minority. Working paper, Northwestern University, Evanston, IL.

Chamley, C. P. 2004. *Rational Herds: Economic Models of Social Learning*. Cambridge University Press, Cambridge, UK.

Debo, L., C. Parlour, U. Rajan. 2007. The value of congestion. Working paper, University of Chicago, Chicago.

Gans, N. 2002. Customer loyalty and supplier quality competition. *Management Sci.* **48** 207–221.

Hassin, R., M. Haviv. 2003. *To Queue or Not To Queue: Equilibrium Behavior in Queuing Systems*. Kluwer Academic Publishers, Norwell, MA.

Hill, J. 2007. Interview: Penn's provost. Penn Current, Philadelphia.

Lariviere, M. A., J. A. van Mieghem. 2004. Strategically seeking service: How competition can generate Poisson arrivals. *Manufacturing Service Oper. Management* **6**(1) 23–40.

Manski, C. 2000. Economic analysis of social interactions. *J. Econom. Perspect.* **14**(3) 115–136.

Neuts, M. F. 1981. *Matrix-Geometric Solutions in Stochastic Models: An Algorithmic Approach*. Dover Publications, New York.

Smith, L., P. Sorensen. 1998. Rational social learning with random sampling. Working paper, University of Michigan, Ann Arbor.

Su, X., S. A. Zenios. 2004. Patient choice in kidney allocation: The role of the queueing discipline. *Manufacturing Service Oper. Management* **6**(4) 280–301.

Su, X., S. A. Zenios. 2005. Patient choice in kidney allocation: A sequential stochastic assignment model. *Oper. Res.* **53**(3) 443–455.

Veeraraghavan, S., L. Debo. 2008. Is it worth the wait? Service choice when waiting is expensive. Working paper, Wharton School, University of Pennsylvania, Philadelphia.

Whinston, W. 1977. The optimality of joining the shortest queue discipline. *J. Appl. Probab.* **14** 181–189.

Whitt, W. 1986. Deciding which queue to join: Some counterexamples. *Oper. Res.* **34**(1) 55–62.

Wolff, R. W. 1982. Poisson arrivals see time averages. *Oper. Res.* **30**(2) 223–231.