Herding in Queues with Waiting costs: Rationality and Regret [1]

Senthil K. Veeraraghavan
Wharton School
University of Pennsylvania
senthilv@wharton.upenn.edu

Laurens G. Debo
Chicago Booth School of Business
University of Chicago
Laurens.Debo@chicagobooth.edu

August 2010

Abstract

We study how consumers with waiting cost disutility choose between two congested services of unknown service value. Consumers observe an imperfect private signal indicating which service facility may provide better service value, as well as the queue lengths at the service facilities before making their choice. If more consumers choose the same service facility because of their private information, longer queues will form at that facility and indicate higher quality. On the other hand, a long queue also implies more waiting time. We characterize the equilibrium queue-joining behavior of arriving consumers, and the extent of their learning from the queue information in the presence of such positive and negative externalities. We find that when the arrival rates are *low*, utility-maximizing rational consumers herd and join the longer queue, ignoring any contrary private information. We show that even when consumers treat queues as independently evolving, herd behavior persists with consumers joining longer queues above a threshold queue difference. However, if the consumers seek to minimize ex-post regret when making their decisions, herd behavior may be dampened.

Keywords: *Herd Behavior, Queueing Games, Learning, Regret, Bounded Rationality.*

1. Introduction

The quality of a service is often difficult to assess when consumers have to choose between two comparable service facilities (such as restaurants). Consumers may have some information about which service is better from previous experiences with similar services, advice from friends or colleagues, etc. This private information alone may not be sufficient for some less-informed consumers to make their decision. With no other information available to them upon arrival to the market, these consumers may be influenced in their decision-making by the length of the queue in front of each service facility. Their decision problem is complicated: On one hand, consumers prefer joining the service facility with the shorter queue as the expected waiting costs will be lower. On the other hand, they may be attracted to the service facility with the longer queue, inferring that

[1]The authors would like to thank Krishnan Anand, Sushil Bikhchandani, Gerard Cachon, Marshall Fisher, Noah Gans, Steve Graves, Serguei Netessine, Alan Scheller-Wolf, Assaf Zeevi and the participants at the MSOM Service Operations SIG Conference, University of Maryland, DOTM Colloquium at UCLA for their comments and suggestions on the paper.

better-informed consumers in the line have 'voted with their feet' and find it worthwhile to wait for the service. The latter, follow-the-crowd behavior has been widely touted in the popular literature as collective wisdom (Surowiecki 2005).

As consumer decisions upon arrival depend on queue lengths, and because queue lengths depend on this decision-making, the arrival rate at each service facility needs to be determined in equilibrium. While many of us have probably faced a situation in which we let our choice between two comparable restaurants depend on the crowd waiting at the restaurants, little is known, about how congestion and delay sensitivity impact consumers' perceived service value in general. Becker (1991) studies a model in which consumers choose between two restaurants based on the congestion levels at those restaurants. Becker's explanation is static and is based on consumption externalities in which consumers derive a higher utility from dining at crowded places. In contrast, our approach explains the decisions of consumers based on learning.

Our paper presents a model of queue choice behavior when consumers face positive information externalities and negative waiting cost externalities. In addition, since some consumers may have more precise information than other consumers, our model explores the impact of heterogeneity in consumers' private information on their decision making. We consider queue joining behavior under different behavioral assumptions: First, we model consumers as rational Bayesian agents. Next, we relax the rationality assumption and consider agents who make specific boundedly rational decisions. Finally, we consider the choice behavior of regret minimizing consumers. For analytical tractability, we assume limited waiting space in front of the service facilities.

Hassin and Haviv (2002) provide a comprehensive survey of the literature dealing with the queue joining behavior of consumers at service facilities in queuing systems. In this literature, consumers are rational decision makers. The papers have focused on minimizing waiting costs; Whinston (1977), for example, shows that when the service process is exponentially distributed at identical servers, joining shorter queues minimizes the expected waiting costs. Further, in this literature, there is no uncertainty about the expected service quality at each server. There is no update based on service choices of previous consumers in the system.

Our paper is related to the research on sequential decisions of rational Bayesian agents who learn from others (analyzed in Section 3). Bikhchandani, Hirshleifer and Welch (1992) and Banerjee (1992) model how informational cascades occur when a series of actors make a decision that is observed by others, in which every subsequent actor, based on the observations of others' actions, makes the same choice independent of her private signal. Chamley (2004) provides a comprehensive

survey on herding literature in economics. Callander and Horner (2009) consider information externalities in a market where consumers are either fully informed or uninformed, with no waiting costs. Veeraraghavan and Debo (2009) focus on how informational externalities emerge due to imperfect signals by analyzing a queue choice model *without* any waiting costs, and without any externalities due to blocking. While the above papers explicitly model the information structure, none of them incorporate negative externalities due to congestion. A working paper by Debo et al. (2007) is an exception. However, they only look at a single-queue model in which the consumers decide whether to join or balk a queue after observing previous decisions. To allow for a study of longer queue joining behavior under waiting cost externalities, we model two service facilities alongside one another, which requires us to solve the underlying finite two-dimensional birth and death processes.

We will show that to determine the equilibrium, rational consumers would need to solve for the stationary probabilities for the underlying 2-dimensional birth and death process, for a multitude of candidate strategies. Given the computational effort required in the characterization of the equilibrium, it is likely that consumers make *boundedly rational* decisions due to cognitive constraints. A stream of literature in economics originating from the seminal paper by Simon (1955) explores bounded rationality in problems where complexity may be an issue. Our queue choice model based on queue independence, is a specific behavioral assumption motivated by the discussions in Rubinstein (1998) on how consumers may adopt simplifying approaches to restricting or manipulating the information they use.

Finally, we also examine consumer decisions under *regret* in congestion-prone environments (in Section 4). Regret theory (Bell 1982, Loomes and Sugden 1982) explains deviations from the expected utility theory for decisions under uncertainty. Regret refers to the ex post comparison between chosen alternative and the optimal alternative. Consumers minimize their anticipated regret or disappointment from not choosing the ex post optimal alternative (the better service facility), instead of maximizing their ex ante expected utility. Thus regret is an ex ante examination of ex post outcomes. For instance, in Schweitzer and Cachon (2000), the decision-makers in newsvendor settings try to minimize regret by choosing quantities that reduce ex post "errors".

The main objective of this paper is to explore how consumer choice and learning emerge in congestion-prone environments. We summarize some of our main observations:

1. We show that the rational Bayesian consumers join the longer queue when the arrival rates are *low* compared to the service rate. This is a somewhat intriguing result. In other words, long

queues are more informative about the service value when service rates are much higher than arrival rates because the waiting space constraints and stochastic departures do not contaminate the extent of learning from other consumers.

2. We show that the equilibrium queue joining strategy of rational Bayesian consumers can be complex. We find that consumers may avoid empty queues, and join a longer queue despite incurring additional waiting costs. When both queues are non-empty, the longer queue joining behavior occurs when one queue is sufficiently more crowded than the other (i.e., the queue-difference is large enough), *and* when the market is sufficiently crowded (i.e., roughly when the *sum* of the queue lengths is high enough). Intriguingly, rational Bayesian consumers may ignore their private information, and join the shorter, non-empty queue (i.e., they buck the trend) when the market is not crowded.

3. If consumers deviate are not Bayesian and treat the queues as two independent entities (i.e., simplifications in calculating the stationary probabilities), we find that longer queue joining behavior is even more pervasive. It persists at all arrival rates. The longer queue joining behavior is of threshold type, i.e., the consumers join the longer queue when the queue difference is above a certain threshold.

4. Finally, if a consumer minimizes the expected ex post regret from choosing a queue, the longer queue joining behavior persists. We find that the expected regret minimizing strategy is identical to the equilibrium strategy of rational Bayesian agents. However, if consumers minimize the worst case regret (the minimax criterion of Savage (1951)), they always join the shorter queue at all states.

2. Model

The game: We consider a game in which consumers arrive sequentially according to a Poisson process with arrival rate λ to a market with two servers. In front of each server, a queue whose length is at most N (including the consumer in service) can be formed.

The service: The exact service value of the two facilities, (V_1, V_2), is the same for all consumers, but unknown. It is the net utility of obtaining the service. Its joint distribution, $F(v_1, v_2)$ over $[\underline{v}, \overline{v}] \times [\underline{v}, \overline{v}]$ with $\underline{v} < \overline{v} \in \mathbb{R}$ is common knowledge. Let $f(v_1, v_2)$ be the density function of the distribution of the valuations. We make no further distributional assumption on $f(\cdot)$ except that it is symmetric and continuous. Service time at both servers is exponentially distributed with mean

τ. Agents incur a waiting cost per unit of time that they are in the system: $c \geq 0$. We define *traffic intensity* as $\rho = \lambda\tau$. The arrival rate can be arbitrarily different compared to the service rate, i.e., $0 < \rho < \infty$.

The consumer information: Upon arrival to the market, all consumers observe the queue length in front of each service facility; $\mathbf{n} = (n_1, n_2) \in \overline{\mathcal{N}} \triangleq \{0, \cdots, N\} \times \{0, \cdots, N\}$. All consumers also receive a private signal $s \in \mathcal{S} = \{1, 2\}$. This signal is an indicator of which service facility provides the highest value in the market.

The market consists of two consumer classes. The class to which a particular consumer belongs is private information to the consumer. Class 1 consumers are perfectly informed about which server provides better value. Let α represent the fraction of such consumers, who are henceforth referred to as fully informed consumers, or simply, *informed consumers*. For informed consumers, $\Pr(s = 1 \mid V_1 > V_2) = \Pr(s = 2 \mid V_1 < V_2) = 1$. Consumers do not know the exact values of (V_1, V_2) but know whether $V_1 > V_2$ or $V_1 < V_2$. Our conclusions can be generalized for markets with multiple classes where the most informed consumers are also imperfectly informed.

The rest of the market is composed of Class 2 consumers or *less-informed consumers*, who receive a signal $s \in \mathcal{S}$ such that: $\Pr(s = 1 \mid V_1 > V_2) = \Pr(s = 2 \mid V_1 < V_2) = g \in [1/2, 1)$, meaning that if the true state is that server i provides a better value than server j, each less-informed consumer receives a signal $s = i$ $(s = j)$ with probability g $(1 - g)$. We will refer to the parameter g as *signal strength*. The classes reflect the fact that some consumers have better information than others. Thus, even though each consumer knows her own signal strength, she does not know the signal strengths of other consumers, and knows only its distribution.

Consider any consumer that arrives at the market. Let $\mathcal{A} = \{0, 1, 2\}$ be the set of possible actions that the consumer can take upon arrival; 1 represents joining server 1, 2 represents joining server 2, and 0 represents the consumer being blocked due to buffer constraints.

Let $V_+ = \mathbb{E}[V_1 | V_1 > V_2] (= \mathbb{E}[V_2 | V_2 > V_1])$ and $V_- = \mathbb{E}[V_2 | V_1 > V_2] (= \mathbb{E}[V_1 | V_2 > V_1])$ be updated valuations of the better and worse service facilities in the market, conditional on one of them being better than the other, and their difference be $\Delta = V_+ - V_-$. We assume no balking, jockeying or reneging in the queues. We focus on the selection between two service providers, and therefore assume the expected valuations are such that $V_- > Nc\tau$. In other words, the expected service value of the low-quality service facility is larger than the expected waiting costs when there are $N - 1$ consumers in the queue. The no-balking condition allows us to ignore strategies that involve balking in the action space of consumers and focus on the key phenomenon of interest: the

equilibrium queue *selection* behavior – when and why consumers join longer queues instead of joining queues that have lower waiting times.

Let $\sigma_k^j(s, \mathbf{n})$ be the strategy of consumer j of class $k \in \{1, 2\}$. $\sigma_k^j(s, \mathbf{n}) = a$ denotes that consumer j of class k joins queue a after observing state $\mathbf{n}$ and signal s. As each server can contain N consumers, all arriving consumers are blocked if there is no waiting space, i.e. $\forall k, s\ \sigma_k^j(s, (N, N)) = 0$. In other words, no waiting consumer is 'bumped' to accommodate another. The consumers differ only in their private information (which is unidentifiable), and not in their service priority. When one queue is full and there is waiting space available in the other queue, consumers join the other queue, even if it provides lower valuation than the blocked queue, since the net utility from the second queue is positive. Therefore, the actions are $\sigma_k^j(s, (n, N)) = 1$ and $\sigma_k^j(s, (N, n)) = 2$ (for $n \in \{0, ..., N-1\}$, $\forall\ k \in \{1, 2\}$, and $\forall j$). These actions represent consumers joining the queue of a competing service facility when their preferred server is full. As a result, we need to determine the equilibrium consumer actions for the set of states $\mathcal{N} \triangleq \{0, \cdots, N-1\} \times \{0, \cdots, N-1\}$. Since all consumers within a class are homogeneous ex ante, we consider symmetric strategies within each class (and allow for varying strategies across different classes); $\sigma_k^j = \sigma_k$ for all consumers j in each class k. We can now formally define herding in our context.

Definition 1. *A consumer of class k "herds" (at state $\mathbf{n}$), when she ignores her signal and joins the longer queue, i.e. given $n_1 > n_2$, $\sigma_k(s, \mathbf{n}) = 1$ for $s \in \{1, 2\}$ and for $n_1 < n_2$, $\sigma_k(s, \mathbf{n}) = 2$ for $s \in \{1, 2\}$*

Hence with our definition, when a consumer herds, she may observe a signal pointing to the shorter queue and yet join the longer queue. First, note that the strategy of the fully informed (Class 1) consumers $\sigma_1(s, \mathbf{n})$ can immediately be completely specified for all of the decision-making criteria that we study in this paper. Since their signal is perfectly informative, they join the queue corresponding to the signal (if there is space) as long as the additional expected utility at the better server is greater than any additional expected waiting cost incurred. For example, if the signal is 1, $\sigma_1(1, \mathbf{n}) = 1$ as long as $V_+ - c(n_1 + 1)\tau > V_- - c(n_2 + 1)\tau$ or $\Delta > c(n_1 - n_2)\tau$. Otherwise, they join the shorter queue. Since the informed consumers' strategy is fully specified, we focus on the less-informed (Class 2) consumers' strategy. Hence, throughout the paper, we suppress the strategies of the informed consumers, and address only the less-informed consumers. For notational convenience, when the less-informed consumers follow the strategy $\sigma_2 = \sigma$, we will use σ to denote the strategies of all consumers in the market. In the following section, we model the strategies of the less-informed consumers as rational Bayesian agents.

3. Rational Bayesian Consumers

Consider a consumer j and fix the strategy of all consumers $j' \neq j$ at σ'. Denote consumer j's belief of the service value upon observing $\mathbf{n}$ as $f(v_1, v_2 \mid \mathbf{n}, \sigma')$. After observing $(\mathbf{n}, s)$, a randomly arriving less-informed consumer updates her prior expected service value for both service facilities:

$$\mathbb{E}\left(V_i \mid \mathbf{n}, s; \sigma'\right) = \int_{\underline{v}}^{\overline{v}} \int_{\underline{v}}^{\overline{v}} v_i f\left(v_1, v_2 \mid \mathbf{n}, s; \sigma'\right) dv_2 dv_1, \ i \in \{1, 2\}.$$

Let $BR(\sigma')$ be the best response of a consumer to some σ'. Then, $\sigma \in BR(\sigma')$ if and only if for $i \in \{1, 2\}$ and all $\mathbf{n} \in \mathcal{N}$:

$$\mathbb{E}\left(V_i \mid \mathbf{n}, s; \sigma'\right) - cn_i\tau > \mathbb{E}\left(V_{-i} \mid \mathbf{n}, s; \sigma'\right) - cn_{-i}\tau \Rightarrow \sigma(s, \mathbf{n}) = i. \tag{1}$$

Now, we can define conditions for a pure strategy Markov Perfect Bayesian equilibrium (Fudenberg and Tirole 1991, Maskin and Tirole 2001):

Definition 2 (Rational Bayesian Consumers)**.** *A strategy σ^* is a pure strategy stationary Markov Perfect Bayesian equilibrium if, $\sigma^* \in BR(\sigma^*)$, and $f(v_1, v_2 \mid \mathbf{n}, s; \sigma^*)$ is defined by Bayes' rule for any $\mathbf{n}$ that is reached on the equilibrium path with a positive probability.*

We are now ready to characterize the equilibrium strategies of all consumers, i.e., we characterize σ^* based on the long-run probability distributions of the queue states. For a given strategy vector 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), and σ representing the consumer's strategy. Using the PASTA property (Wolff 1982), $\pi_i(\mathbf{n}, \sigma)$ is also the probability that any randomly arriving consumer sees state $\mathbf{n}$, conditional on $V_i > V_{-i}$. From Bayes' Theorem, the updated density of the service value is:

$$f(v_1, v_2 \mid \mathbf{n}, s; \sigma) = \begin{cases} \frac{g(s)}{D(\mathbf{n}, s, \sigma)} \pi_i(\mathbf{n}, \sigma) f(v_1, v_2) & v_1 > v_2 \\ \frac{1-g(s)}{D(\mathbf{n}, s, \sigma)} \pi_{-i}(\mathbf{n}, \sigma) f(v_1, v_2) & o/w \end{cases} \tag{2}$$

where $g(1) = g$, $g(2) = 1 - g$, and $D(\mathbf{n}, s, \sigma)$ is a normalization constant such that

$$D(\mathbf{n}, s, \sigma) = g(s) \pi_i(\mathbf{n}, \sigma) \int_{\underline{v}}^{\overline{v}} \int_{\underline{v}}^{v_1} f(v_1, v_2) dv_2 dv_1 + (1 - g(s)) \pi_{-i}(\mathbf{n}, \sigma) \int_{\underline{v}}^{\overline{v}} \int_{v_1}^{\overline{v}} f(v_1, v_2) dv_2 dv_1.$$

Given the consumer actions at each state in the system, the probabilities of the system being in the state $\pi_i(\mathbf{n};\sigma)$ can be derived. Let $l(\mathbf{n}; \sigma) \triangleq \frac{\pi_1(\mathbf{n};\sigma)}{\pi_2(\mathbf{n};\sigma)}$ denote the ratio of the probability of seeing

state $\mathbf{n}$ when server 1 is better than server 2 to the probability of seeing state $\mathbf{n}$ when server 2 is better than server 1. In each state $\mathbf{n} \in \mathcal{N}$, an equilibrium action pair needs to be determined for each consumer.

Lemma 1. *For a given $l(\mathbf{n};\sigma)$, assume that $n_1 > n_2$, then for a less-informed consumer:*

$$\begin{cases} \sigma(s,\mathbf{n}) = 1 & \Leftrightarrow & \frac{g}{1-g}\frac{\Delta+(n_1-n_2)c\tau}{\Delta-(n_1-n_2)c\tau} & < & l(\mathbf{n};\sigma) & & & (L) \\ \sigma(s,\mathbf{n}) = s & \Leftrightarrow & \frac{1-g}{g}\frac{\Delta+(n_1-n_2)c\tau}{\Delta-(n_1-n_2)c\tau} & < & l(\mathbf{n};\sigma) & < & \frac{g}{1-g}\frac{\Delta+(n_1-n_2)c\tau}{\Delta-(n_1-n_2)c\tau} & (F) \\ \sigma(s,\mathbf{n}) = 2 & \Leftrightarrow & & & l(\mathbf{n};\sigma) & < & \frac{1-g}{g}\frac{\Delta+(n_1-n_2)c\tau}{\Delta-(n_1-n_2)c\tau} & (S) \end{cases}$$

Lemma 1 summarizes the conditions for the equilibrium actions at a state, given the signal strength and waiting costs. When the first condition is satisfied, the consumer *herds* (See Definition 1; even if the private signal points to the shortest queue, i.e., $s = 2$, the consumer will join the longer queue, i.e., $\sigma(s,\mathbf{n}) = 1$.). In the middle condition, the consumer follows her signal, and when the third condition is satisfied, the consumer follows the shorter queue irrespective of her signal. In what follows, we examine the conditions under which a consumer will herd. Consider a consumer arriving at $\mathbf{n}$ (with $n_1 > n_2$). Suppose all arriving consumers follow some strategy σ.

If server 1 is better, the observation $\mathbf{n}$ occurs with probability $\pi_1(\mathbf{n};\sigma)$. *(i)* With probability g, the consumer observes the correct signal 1. Since the signal points to the longer queue, there is no additional utility in ignoring the signal and following the longer queue. *(ii)* On the other hand, she could see an 'incorrect' signal with probability $1-g$. The additional marginal utility from herding is $(V_+ - c(n_1+1)\tau) - (V_- - (n_2+1)c\tau)$. Therefore, the ex-ante expected additional utility from ignoring the signal and joining the longer queue at state $\mathbf{n}$ is $\pi_1(\mathbf{n};\sigma)(1-g)(\Delta - c(n_1-n_2)\tau)$.

In the alternate case server 2 is better (i.e., the longer queue is at inferior server). Then, the observation of state $\mathbf{n}$ occurs with probability $\pi_2(\mathbf{n};\sigma)$. *(i)* The consumer sees incorrect signal 1 (same as the longer queue) with probability $(1-g)$. Therefore, there is no additional disutility in joining the longer queue (over following the signal). *(ii)* However, with probability g, the consumer sees a correct signal 2. By ignoring the signal and joining the longer queue, she will suffer an additional disutility of $V_+ - c(n_2+1)\tau - (V_- - c(n_1+1)\tau)$. Therefore, the expected additional disutility from joining the longer queue at $\mathbf{n}$ is $g\pi_2(\mathbf{n};\sigma)(\Delta + c(n_1-n_2)\tau)$.

A rational consumer always joins the longer queue when $\pi_1(\mathbf{n};\sigma)(1-g)(\Delta - c(n_1-n_2)\tau) > \pi_2(\mathbf{n};\sigma)g(\Delta + c(n_1-n_2)\tau)$ or simply when, $l(\mathbf{n};\sigma) > \frac{g}{1-g}\frac{\Delta+c(n_1-n_2)\tau}{\Delta-c(n_1-n_2)\tau}$. Equipped with Lemma 1, we obtain the equilibrium strategy when the queues are equal in length.

Corollary 2. *We have $\sigma^*(s,(n,n)) = s$ for all $n \in \{0,...,N-1\}$: when the queue lengths are identical, the equilibrium action is to follow one's private signal.*

Corollary 2 states that consumers follow their private signals when they observe that the queue lengths are equal (i.e. $n_1 = n_2$). This is intuitive: a less-informed consumer obtains no additional information about relative service value when queue lengths are equal. In general, characterizing an equilibrium strategy analytically is very difficult. We will begin our analysis of rational consumers by considering queues with small buffers in sections §3.1 and §3.2, and examine the behavior of consumers when buffers are large in §3.3.

3.1 Analysis: the case of small queue buffers $N = 2$

In this section, we limit the waiting space in front of each of the service facilities to two (i.e. $N = 2$). When one queue is full, the consumers join the other queue and at $(2,2)$, they are blocked from joining. From Corollary 2, the consumers follow their signal at states (n,n) with $0 \leq n \leq 1$. As a result, the only actions that need to be specified in equilibrium are at states $(1,0)$ and $(0,1)$. Since the servers are symmetric and are indistinguishable ex ante, the equilibrium action will also be symmetric at states $(1,0)$ and $(0,1)$. We thus focus on the only state for which the equilibrium action needs to be determined, and introduce the following notation: $\sigma^{\mathbf{F}}$ indicates that $\sigma(s,(1,0)) = s,\ \forall s \in \{1,2\}$, i.e., consumers follow their signal; $\sigma^{\mathbf{S}}$ indicates that $\sigma(s,(1,0)) = 2\ \forall s \in \{1,2\}$, i.e., consumers always join the shorter queue (ignoring their private signal); and $\sigma^{\mathbf{L}}$ indicates that $\sigma(s,(1,0)) = 1\ \forall s \in \{1,2\}$, i.e. consumers join the longer queue, ignoring their private signal (i.e., consumers *herd*). We focus on $\Delta > c\tau$ since when $\Delta < c\tau$, consumers always join the shorter queue (See Lemma C1 in Appendix C.). Proposition 3 provides a special case when the less-informed consumers are fully uninformed (i.e. $g = 1/2$).

Proposition 3. *The equilibrium strategy for the uninformed consumers ($g = 1/2$) is as follows.*

1. *$\sigma^{\mathbf{S}}$ is an equilibrium (i) $\forall \rho > 0$ when $\alpha\Delta < c\tau < \Delta$, (ii) $\forall \rho \in (\hat{\rho}, \infty)$ when $c\tau \leq \alpha\Delta$.*
2. *$\sigma^{\mathbf{L}}$ is an equilibrium (i) $\forall \rho \in (\underline{\rho}, \overline{\rho})$ when $\alpha\Delta < c\tau < \hat{\alpha}\Delta$, (where $\alpha < \hat{\alpha}$) (ii) $\forall \rho \in (0, \overline{\rho}]$ when $c\tau \leq \alpha\Delta$.*

For sufficiently low waiting costs, i.e. when $c\tau < \alpha\Delta$ (where α denotes the fraction of informed customers), the less-informed consumers herd (i.e., $\sigma^{\mathbf{L}}$ is an equilibrium) when the traffic intensity is lower than a threshold, $\overline{\rho}$ (see Proposition 3.2(ii)). When the waiting cost is higher (i.e., $\alpha\Delta <$

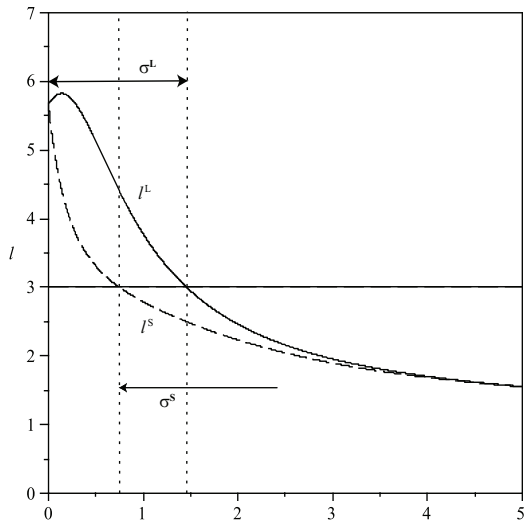

Figure 1: *Likelihood ratios l^L, l^S (under strategies $\sigma^{\mathbf{S}}$ and $\sigma^{\mathbf{L}}$ respectively) are plotted again ρ on the x-axis using thick and dotted curves respectively. The threshold is indicated by horizontal line. Applying Lemma 1, the equilibrium strategies are shown by arrow marks. For instance, Lemma 1 requires l^L to be higher than the threshold. This occurs for ρ less than the right-most vertical dotted line which denotes ($\rho = \overline{\rho}$). Thus, the joining the longer queue $\sigma^{\mathbf{L}}$ is in equilibrium for low arrival rates (i.e. for $\rho \in (0, \overline{\rho}]$).*

$c\tau < \hat{\alpha}\Delta$), they herd at intermediate traffic intensities. As the traffic intensity increases, consumers are more likely to see full buffers, but are also more likely to rationalize that consumers joined a certain queue either because the other queue was blocked, or because it was too costly to wait at the alternative queue. In Figure 1, we illustrate the region of longer-queue joining by plotting the likelihood ratios at $(1, 0)$, (i.e, l^S and l^L under strategies $\sigma^{\mathbf{S}}$ and $\sigma^{\mathbf{L}}$ respectively), as a function of λ and $\frac{\Delta + c\tau}{\Delta - c\tau}$.

Extent of Learning: To measure the extent of learning, we first examine the expected valuation of a server to consumers in a system that does not make queue lengths publicly available. Based on her private signal s, the expected value from a server $i \in \{1, 2\}$ is $\mathbb{E}(V_i|s)\ \forall s \in \{1, 2\}$. Since the informed consumers are perfectly informed, we focus on uninformed consumers (i.e. $g = 1/2$, as in Proposition 3). For the uninformed consumers, based on the private information alone, the expected valuation from either server for the uninformed consumer is $(V_+ + V_-)/2$ (regardless of their private signal). Suppose that the queue lengths are made available to the consumer on arrival. We focus on the state $(1, 0)$. Up on arrival, the expected waiting costs are identical for both informed and uninformed consumers (in this case, $2c\tau$ at the longer queue). We now characterize how much the expected valuation from the longer queue changes for an uninformed consumer. Let the change in expected valuation from the prior valuation, for a consumer with signal s at queue i, be denoted by

$V_q(i, s, \mathbf{n}, \sigma)$, when all consumers play a strategy σ. Hence, $V_q(i, s, \mathbf{n}, \sigma) = \mathbb{E}(V_i|\mathbf{n}, s, \sigma) - (V_+ + V_-)/2$ for uninformed consumers with $g = 1/2$. In Proposition 4, we characterize the extent of learning due to the revelation of queue length information by specifying $V_q(1, s, \mathbf{n}, \sigma)$.

Proposition 4. *Given consumer strategy σ at state $(1,0)$, $V_q(1, s, (1,0), \sigma) = \left(\frac{l(1,0;\sigma)-1}{l(1,0;\sigma)+1}\right)\frac{\Delta}{2}$, $\forall\ s \in \{1,2\}$. For any $\alpha > 0$, $\exists \overline{\rho} > 0$ such that $\forall \rho \in [0, \overline{\rho}]$ $V_q(1, s, (1,0), \sigma^{\mathbf{L}}) \geq \alpha\Delta/2 = \max_{\rho \geq 0}\{V_q(1, s, (1,0), \sigma^{\mathbf{S}})\}, \forall s \in \{1,2\}$.*

If consumers follow $\sigma^{\mathbf{S}}$ the maximal additional valuation is bounded at $\alpha\Delta/2$. However Proposition 4 shows that when consumers herd, the queue lengths are more informative at low traffic intensities. Since the additional information from the queue lengths is beneficial for everyone, this causes consumers to play $\sigma^{\mathbf{L}}$ in equilibrium at low traffic intensities.

Now we expand the analysis from the special case in Proposition 3 to the case when the less-informed consumers have a signal strength $g \in [1/2, 1)$. It can be seen that informed consumers will always follow their signal (also see discussion following Definition 1). The less-informed consumers adopt an equilibrium strategy characterized by the conditions in Lemma 1. Proposition 5 sheds more light on the market conditions when the less-informed consumers herd and the extent of learning when they herd.

Proposition 5. *There exist $\overline{c}$ and $\overline{\rho}$, such that $\forall\ \rho < \overline{\rho}$ and $c < \overline{c}$, for the less-informed consumers, (i) $\sigma^{\mathbf{L}}$ is in equilibrium, (ii) $V_q(1, 2, (1,0), \sigma^{\mathbf{L}}) = \Delta\frac{g(1-g)(l^L-1)}{(g+l^L(1-g))} \geq V_q(1, 1, (1,0), \sigma^{\mathbf{L}}) = \Delta\frac{g(1-g)(l^L-1)}{(gl^L+(1-g))} \geq \max_{\rho \in [0,\overline{\rho}]}\{V_q(1, 1, (1,0), \sigma^{\mathbf{L}}), V_q(1, 1, (1,0), \sigma^{\mathbf{S}})\}$.*

Proposition 5(i) generalizes the findings about the relationship between herding and arrival rates made in Proposition 3, which was only valid for uninformed consumers (i.e., $g = 1/2$). In equilibrium, the less-informed consumers, herd at low arrival rates for sufficiently low waiting costs, and they expect an increased value from observing queue information, especially when they see an opposite signal to a longer queue. The results from Proposition 4 are generalized for $g > 1/2$ in Proposition 5(ii), i.e., when consumers herd, the additional valuation gained from the longer queue is high at low arrival rates.

3.2 Analysis: the case of small queue buffers $N = 3$

We increase buffer size to $N = 3$ in this section, and note that $\sigma(s, (n, 3)) = 1$ and $\sigma(s, (3, n)) = 2$ for $n \in \{0, 1, 2\}$: When one queue is full, the consumers join the other queue. At $(3, 3)$ they are all blocked. From Corollary 2, consumers follow their signal when queue lengths are equal.

As a result, we examine states: $\{(1,0),(2,1),(2,0)\}$ and $\{(0,1),(1,2),(0,2)\}$. Since we focus on symmetric equilibrium strategies, we can limit our attention to the three states $(1,0),(2,1),(2,0)$. As a shortcut in notation, we indicate an equilibrium strategy by $\sigma^{\mathbf{XYZ}}$, with $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Z}$ $\in\{\mathbf{F},\mathbf{L},\mathbf{S}\}$ being the strategy at $\{(1,0),(2,1),(2,0)\}$ (and symmetrically at $\{(0,1),(1,2),(0,2)\}$), respectively. The equilibrium strategies are specified by Lemma 1.

We rewrite the likelihood ratio conditions established in Lemma 1 for each of the 27 possible pure strategy equilibria as a function of ρ, such that:

$$\left\{\begin{array}{ccccc} \bar{l}\,(1) & < & l^{\mathbf{LYZ}}_{(1,0)} & & \\ \underline{l}\,(1) & < & l^{\mathbf{FYZ}}_{(1,0)} & < & \bar{l}\,(1) \\ & & l^{\mathbf{SYZ}}_{(1,0)} & < & \underline{l}\,(1) \end{array}\right. , \left\{\begin{array}{ccccc} \bar{l}\,(1) & < & l^{\mathbf{XLZ}}_{(2,1)} & & \\ \underline{l}\,(1) & < & l^{\mathbf{XFZ}}_{(2,1)} & < & \bar{l}\,(1) \\ & & l^{\mathbf{XSZ}}_{(2,1)} & < & \underline{l}\,(1) \end{array}\right. \text{ and } \left\{\begin{array}{ccccc} \bar{l}\,(2) & < & l^{\mathbf{XYL}}_{(2,0)} & & \\ \underline{l}\,(2) & < & l^{\mathbf{XYF}}_{(2,0)} & < & \bar{l}\,(2) \\ & & l^{\mathbf{XYS}}_{(2,0)} & < & \underline{l}\,(2) \end{array}\right. \tag{3}$$

with $l^{\mathbf{XYZ}}_{\mathbf{n}} = l\left(\mathbf{n};\sigma^{\mathbf{XYZ}}\right)$, $\bar{l}\,(n) = \frac{g}{1-g}\frac{\Delta+nc\tau}{\Delta-nc\tau}$ and $\underline{l}\,(n) = \frac{1-g}{g}\frac{\Delta+nc\tau}{\Delta-nc\tau}$. We can see that $\sigma^{\mathbf{XYZ}}$ is an equilibrium strategy if the conditions in Equation (3) are satisfied. In states $(1,0)$ and $(2,1)$, the difference in queue lengths is 1, hence, under Lemma 1, $\bar{l}\,(1)$ and $\underline{l}\,(1)$ determine the equilibrium condition. $\bar{l}\,(2)$ and $\underline{l}\,(2)$ determine the equilibrium condition in $(2,0)$.

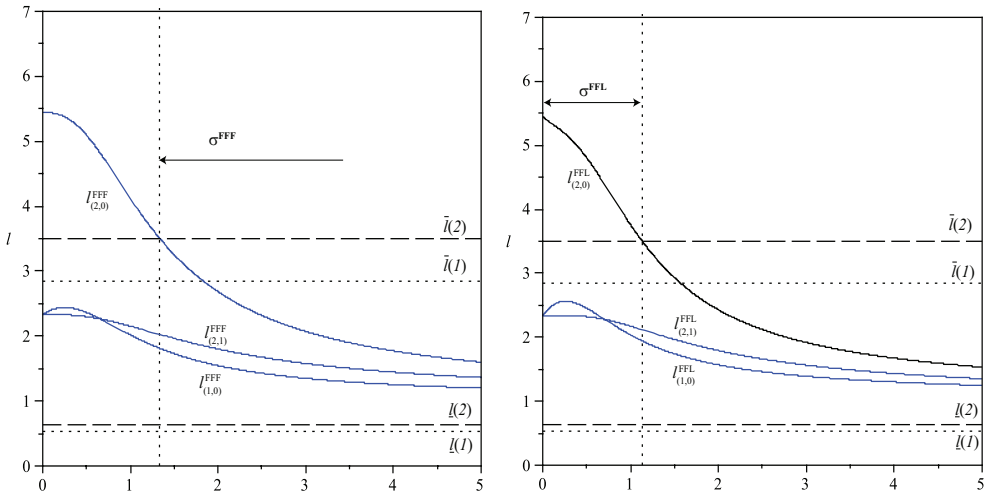

Figure 2: *The existence of equilibrium strategies* $\sigma^{\mathbf{FFF}}$ *(left panel) and* $\sigma^{\mathbf{FFL}}$ *(right panel) over* ρ *(x-axis) are shown for* $g=0.7$ *and* $c=0.1$. *In each subplot, the likelihood ratio functions at three states* $(1,0),(2,1),(2,0)$ *are shown. In the right plot, we note that* $l^{FFL}_{(2,1)}$ *and* $l^{FFL}_{(1,0)}$ *are less than* $\bar{l}(1)$ *for all traffic intensities. They satisfy 'Follow the queue' conditions for left and middle columns in Equation (3) and thus consumers follow their signals at* $(1,0)$ *and* $(2,1)$ *for all* ρ. *Finally,* $l^{FFL}_{(2,0)} > \bar{l}(2)$ *for* ρ *left of the vertical line, therefore customer follow the longer queue at* $(2,0)$. *Therefore, for small traffic intensities,* $\sigma^{\mathbf{FFL}}$ *is in equilibrium. Thus customer herding occurs at traffic intensities bounded by arrow marks in the right panel.*

In Figure 2, we address the strategies σ^{FFF} (left panel) and σ^{FFL} (right panel). We focus on

the longer queue joining behavior at state $(2,0)$ and find that herding occurs at low arrival rates. Herding is also more pronounced at $(2,0)$ than at $(2,1)$ or $(1,0)$ despite the higher waiting costs. The consumer avoids the empty queue and joins the longer queue even if her private signal points to the empty queue. In section 3.3, we will see that herd behavior persists for larger state spaces. We extend the analysis to non-linear waiting costs and asymmetric buffers in §3.4.

3.3 Analysis for Larger Queue Buffers

In this section, we show that herd behavior continues to occur at low traffic intensity for large buffer sizes. Determining an equilibrium strategy for a large state space is a complex problem requiring both the determination of equilibrium strategies of consumers in the queue *and* the calculation of long-run probabilities at each state. In normal-form games, showing the existence of a Nash equilibrium with a specific structure (such as *herding* at some state) is an NP-complete problem (Gilboa and Zemel, 1989). To characterize the equilibrium for our problem, we also have to calculate stationary probabilities for a multitude of candidate strategies. However, the expressions for steady state probabilities of 2-D birth and death process are not known even under specific strategies such as join-the-shortest-queue at all states (Halfin 1985). Steady state probabilities of quasi birth and death processes are of non-product form, and can only be calculated using numerical approaches such as matrix-geometric methods (Neuts 1981). Therefore, we determine an equilibrium strategy profile (within some precision limits) by means of an iterative process that utilizes a regenerative theory based approach (Grassmann et al. 1985). This iterative procedure and related computational issues are described in Appendix A.

For the sake of brevity, we present a representative set of computations, and summarize our observations. In Figures 3 and 4, we show the equilibrium results for $N = 20$, and indicate the states at which the less-informed consumers always join the longer queue (i.e., herd) by L, join the shorter queue by S and follow their signal by F. Given $N = 20$, there are 441 possible states out of which equilibrium actions need to be determined in 380 states (excluding diagonal and upper boundaries). In each state, consumers may either follow their signal, herd, or join the shortest queue. This gives rise to $3^{380} \simeq 10^{181}$ possible equilibrium profiles. As we move from Figure 3 to Figure 4, we increase the fraction of informed consumers in the market α. Within each figure, we first increase λ, and then increase strength of the signal, g.

Based on the analytical insights of the previous sections, we examine the state dependent strategies adopted in equilibrium: (i) As the signals get stronger, the consumers follow their pri-

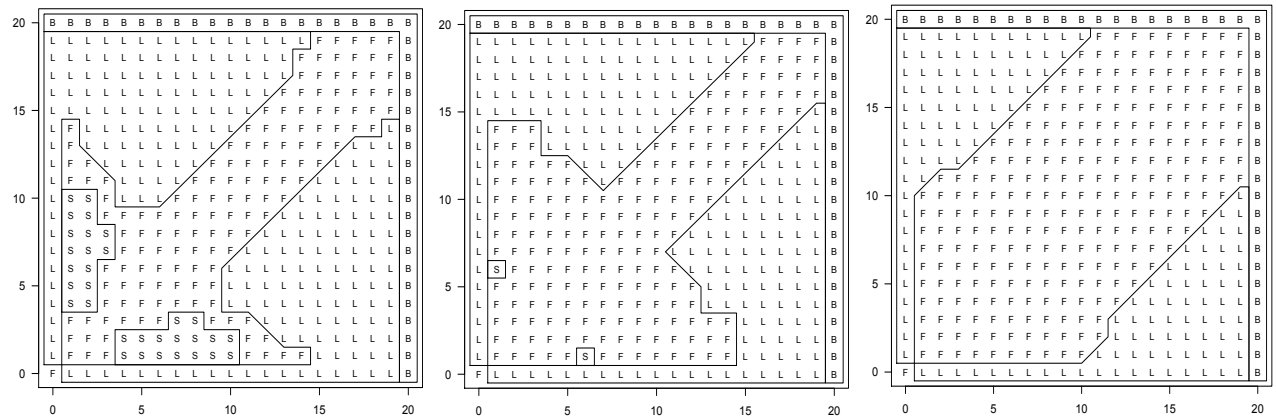

Figure 3: *The equilibrium actions when $N = 20$. In all figures, x- and y- axes are queue lengths of queue 1 and 2 respectively. For all the figures $\alpha = 0.25$, $V_+ = 14$, $V_- = 4$, $\Delta = 10$ and $c = 0.15$. As we move from the left to the middle, ρ increases. Then, moving from the middle figure to the right, g increases. In the figures (a) $\rho = 0.70, g = 0.75$, (b) $\rho = 0.90, g = 0.75$ and (c), $\rho = 0.90, g = 0.99$. At each state in each plot, we denote the longer queue joining (herding) strategy by L, shorter queue joining strategy by S, and following one's signal by F. We denote the consumer being blocked from one (or both) of the queues by B. Observe the dagger-shaped equilibrium queue joining strategy. Note that the longer queue joining occurs when the difference between the queue lengths and the sum of the queue lengths both exceed some thresholds.*

vate signals at more states, evident as we move from the left to the right panel in both figures. (ii) Following Proposition 5, as the fraction of informed consumers α increases, the less-informed consumers herd at more states (observed by comparing the corresponding panels in Figures 3 and 4). (iii) Proposition 5(i) continues to hold. The herd behavior is more pronounced when the arrival rates are low, evident from comparing the two adjacent panels on the left in Figures 3 and 4. (iv) Finally, when one queue is empty, the consumers join the other non-empty queue (as in Becker (1991), Veeraraghavan and Debo (2009)), even though they incur higher waiting costs, i.e., herd behavior persists. In fact at $(8, 0)$, in Figure 3(a), applying the learning results from Section 3.1, we find that the updated value of the longer queue reaches 13.98 which is slightly less than $V_+(= 14)$ (i.e., the conditional value of the better server is almost fully learned at the observation of the state $(8, 0)$). Hence, empty queues persist.

Notice the 'dagger-like' structure of the equilibrium in Figure 3 on the left and the middle panels: not only do consumers follow their private information when the queue lengths are comparable (as is intuitive from Lemma 1), but they also follow their private information at intermediate crowd size i.e., roughly when the *sum* of the queue lengths is not too large or small.

Consider state $(5, 1)$ in Figure 3 (left panel). When the crowd in the market (i.e. the total number of consumers in the market) is low, the less-informed consumers always follow the *shorter*

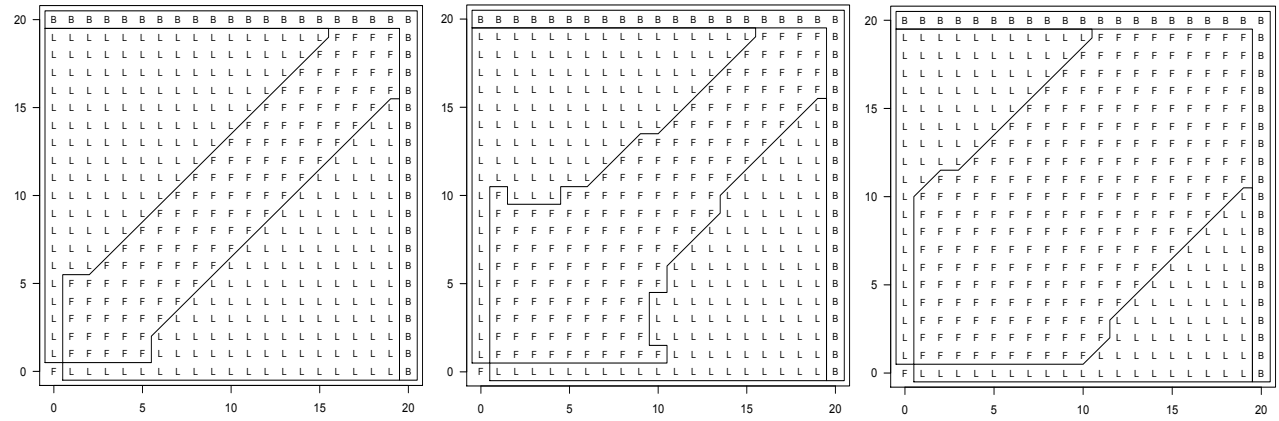

Figure 4: *The equilibrium actions when* $N = 20$*. In all figures, x- and y- axes are queue lengths of queue* 1 *and* 2 *respectively. The scheme and notations are the same as in Figure 3, except that the fraction of informed consumers is higher in the market in this Figure. Specifically:* $\alpha = 0.60$*,* $\Delta = 10$ *and* $c = 0.15$*. For figures (a)* $\rho = 0.70, g = 0.75$ *, (b)* $\rho = 0.90, g = 0.75$*, and (c)* $\rho = 0.90, g = 0.99$*.*

queue. This is not just due to lower waiting costs. Suppose a less-informed consumer observes signal 1. If she were to use only her private signal in updating the service value, she would join the longer queue, since $\mathbb{E}[V_1|1] - (n_1 + 1)c\tau - (\mathbb{E}[V_2|1] - (n_2 + 1)c\tau) = (2g - 1)\Delta - c\tau(n_1 - n_2)$. Suppose the consumer saw signal 1; In our numerical example, the expected values from server 1 and server 2 (based on private signals alone) are 11.5 and 6.5 respectively. If the consumer used queue length information *only* to process waiting costs, the net values at the servers 1 and 2 are $10.6 (= 11.5 - 6 \times 0.15 \times 1)$ and $6.2 (= 6.5 - 2 \times 0.15 \times 1)$ respectively (using $c = 0.15$, $\tau = 1$). Therefore, based on private signals and waiting costs alone, the consumer would join the longer queue.

However, observe that the less-informed consumers join the shorter queue (regardless of their signal). Based on the extent of learning results in §3.1, we can calculate the additional valuation gained on the availability of queue length information. Based on observation of state (5,1), for a consumer who observed signal 1, the $V_q(1, 1, (5, 1), \sigma^*) = -4.31$ and $V_q(2, 1, (5, 1), \sigma^*) = +4.31$ where σ^* is the equilibrium strategy. As a result, the expected values from servers 1 and 2, change to 7.19 and 10.81 respectively. Thus, the supposedly informed minority in the shorter queue is considered more informative than other consumers in the longer queue. Observe from the figure that the less-informed consumer never joins the empty queue instead of a non-empty queue, however the informed consumer will always join the better queue. Thus at (5,1), given the small crowd in the market, there is a significantly increased likelihood of an informed consumer being present at the shorter queue, with no informed consumers at the longer queue. As a consequence, the less-

informed consumer could join the shorter queue. She ignores her signal and joins the shorter queue due to the updating derived from queue lengths. This finding, that rational Bayesian consumers 'buck the trend,' is surprising.

Consider now the state $(18, 1)$ in Figure 3 (left panel). For a less-informed consumer who observes the signal 2, the $V_q(1, 2, (18, 1), \sigma^*) = +6.57$ and $V_q(2, 2, (18, 1), \sigma^*) = -6.57$. This is because when the crowd in the market is large, the conditional probability of having a few informed consumers in the market is high. Further, the crowd is asymmetrically distributed between the facilities, the longer queue is more likely to have those perfectly informed consumers. At that state, the valuation change due to the queue information causes the less-informed consumers to overcome their private information and herd. These actions make the joining strategy pattern look like a dagger.

To conclude, herd behavior is more pronounced at lower traffic intensities than at higher traffic intensities, even for large buffer sizes. This generalizes the findings from Section §3.1. We note that herd behavior is also dependent on the crowd: we observe that herd behavior may not occur in small crowds that are asymmetrically distributed. Instead, the less-informed consumers buck the trend and follow the minority. In the following section §3.4, we show that our findings hold for non-linear waiting costs and asymmetric buffer sizes.

3.4 Extensions: Non-linear Waiting Costs and Asymmetric Buffers

We first extend our findings to non-linear waiting costs. Second, we analyze asymmetric large buffer sizes, and show that our findings continue to hold.

Non-linear Waiting Costs: We consider non-linear waiting costs, specifically a quadratic cost of type $c(n\tau)^2$ where n is the queue length. We note that the conditions for the rational strategies described in Lemma 1 remain identical except that $(n_1 - n_2)c\tau$ is modified to $c((n_1+1)^2 - (n_2+1)^2)\tau^2$.

For small buffers, we observe that the analysis under *non-linear* waiting costs lead to the same conclusions. For $N = 3$, under non-linear waiting costs, the likelihood ratio conditions in Equation (3) becomes

$$\left\{\begin{array}{lcl} \bar{l}\,(3) & < & l^{\mathbf{LYZ}}_{(1,0)} \\ \underline{l}\,(3) & < & l^{\mathbf{FYZ}}_{(1,0)} \\ & & l^{\mathbf{SYZ}}_{(1,0)} \end{array}\right. < \bar{l}\,(3) \quad , \quad \left\{\begin{array}{lcl} \bar{l}\,(5) & < & l^{\mathbf{XLZ}}_{(2,1)} \\ \underline{l}\,(5) & < & l^{\mathbf{XFZ}}_{(2,1)} \\ & & l^{\mathbf{XSZ}}_{(2,1)} \end{array}\right. < \bar{l}\,(5) \quad \text{and} \quad \left\{\begin{array}{lcl} \bar{l}\,(8) & < & l^{\mathbf{XYL}}_{(2,0)} \\ \underline{l}\,(8) & < & l^{\mathbf{XYF}}_{(2,0)} \\ & & l^{\mathbf{XYS}}_{(2,0)} \end{array}\right. < \bar{l}\,(8) \quad (4)$$

For instance, when we consider herding at $(2,0)$ under non-linear costs, $\underline{l}\,(2)$ and $\bar{l}\,(2)$ in the linear cost case increase to $\underline{l}\,(8)$ and $\bar{l}\,(8)$ respectively (since $n_1 - n_2 = 2$ becomes $(n_1+1)^2 - (n_2+1)^2 = 8$). This results in an upward shift of the horizontal line in Figure 2. Again, herd behavior occurs at low arrival rates.

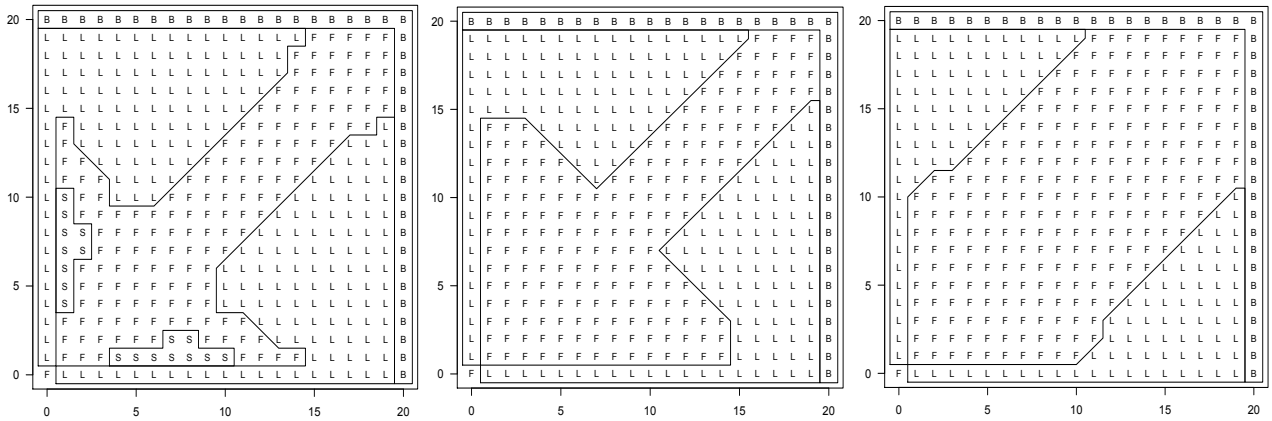

Figure 5: *The equilibrium actions under nonlinear costs when $N = 20$. The parameters are the same as Figure 3.*

The same conclusions hold for higher buffer sizes with non-linear costs. In Figure 5, we report experiments with non-linear waiting costs. Much of the queue choice observed for linear costs continues to hold (herding at low arrival rates, 'dagger'-type equilibrium queue joining, etc.). Higher wait costs cause some additional effects. Consumers may not herd at some states (in which they chose the longer queue under linear waiting costs) since the quality information from the queue lengths may not be sufficient to overcome the higher wait cost. Therefore, we see consumers following their signals in more states. Since the likelihood functions are non-monotone (as seen in Figure 2), we observe the same non-threshold queue choice behavior that we described earlier.

Asymmetric Buffer Sizes: We show that herd behavior is observed for asymmetric buffer sizes. First, we consider small buffer sizes to derive analytical results (similar to the §3.1) by examining a small buffer ($N_1 = 2, N_2 = 1$). For the sake of brevity, the results and the discussion are relegated to Technical Appendix D. Again, we find that consumer herd behavior exists at low arrival rates for sufficiently low waiting costs.

We extend our analysis to large asymmetric buffers (reported in Figure 6). Observe that (i) herding persists at low arrival rates, (ii) an asymmetric 'dagger'-type queue joining behavior is present (i.e. herd behavior depends on the crowd in the market, and how this crowd is distributed between the facilities). Finally, we find that queue joining behavior is largely asymmetric: consumers may not follow their signal when queue lengths are equal.

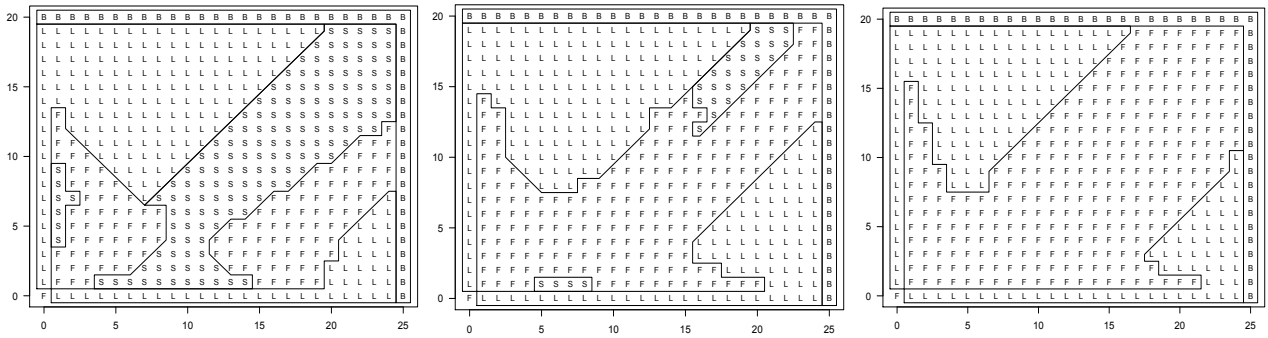

Figure 6: *The equilibrium actions when buffer sizes* $N_1 = 25$ *and* $N_2 = 20$*. In all figures, x- and y- axes are queue lengths of queue* 1 *and* 2 *respectively. The scheme and parameters are the same as Figure 3. The longer queue and the shorter queue joining actions are indicated by L and S respectively.*

3.5 Decision Making: Independent Queues

In our model with rational Bayesian consumers, the main difficulty that consumers face is one of calculating best responses over a precisely imputed steady state distribution, with finite cognitive resources. The closed form solutions for steady state probabilities do not exist even for specific cases such as, join-the-shortest-queue at all states (Kingman 1961, Halfin 1985). Showing the existence of a Nash equilibrium with a specific structure (such as *herding* at some state) is an NP-complete problem (Gilboa and Zemel, 1989). The hardness result suggests real consumers may face significant difficulty in calculating the equilibrium. Thus, we analyze a context where the agents may deviate from rational decisions because of the complexity of their decision problem. Motivated by above reasons, we study the equilibrium outcome of a specific behavioral assumption made by boundedly rational consumers in our queue choice context. (For detailed development of bounded rationality, please refer to Rubinstein (1998) or the quantal choice framework propounded by Luce (1959)).

Our specific behavioral assumption is as follows: as computing the joint probability distribution of the two-dimensional queuing system is a complex and critical task that rational Bayesian consumers face, we relax the rationality assumption by letting the consumers calculate the joint probabilities as being created by two independent queues with finite buffers. Specifically, our consumers derive the steady state probabilities by treating the queues as two independent $M/M/1/N$ queues, *assuming* that all other consumers follow their signals. The service process and buffer sizes of these two queues are common knowledge. Recall that the strategies of the informed consumers are fully specified and are independent of the steady state probabilities. We can thus explore some specific strategies of the less-informed consumers.

Let $q_{i,j}(n,\sigma)$ be the probability that queue i is at state $n \in \{0,1,\ldots,N\}$ conditional on $V_j > V_{-j}$ when such consumers adopt strategy σ. For arrival rates, we impose that the consumers calculate the steady state of a single queue as if all other consumers follow their private information (which we denote by σ'). Thus, the joint probability distribution becomes: $\pi_i\left((n_1,n_2),\sigma'\right) = q_{i,i}(n_1,\sigma') \cdot q_{-i,i}(n_2,\sigma') \forall\, i = 1,2$. All other specifics of the model remain identical to the discussion in Section 2. The updated density function (described in Equation 2) and expected valuation change to,

$$\begin{aligned} f'\left(v_1,v_2 \mid \mathbf{n},s;\sigma'\right) &= \begin{cases} \frac{g(s)}{D(\mathbf{n},s,\sigma)} q_{i,i}(n_1,\sigma') \cdot q_{-i,i}(n_2,\sigma') f\left(v_1,v_2\right) & v_i > v_{-i} \\ \frac{1-g(s)}{D(\mathbf{n},s,\sigma)} q_{i,-i}(n_1,\sigma') \cdot q_{-i,-i}(n_2,\sigma') f\left(v_1,v_2\right) & o/w \end{cases} \\ \mathbb{E}'\left(V_i \mid \mathbf{n},s;\sigma'\right) &= \int_{\underline{v}}^{\overline{v}} \int_{\underline{v}}^{\overline{v}} v_i f'\left(v_1,v_2 \mid \mathbf{n},s;\sigma'\right) dv_2 dv_1,\ i \in \{1,2\} \end{aligned}$$

The best response strategies σ^b for consumers deviating from rational behavior can be identified by applying Equation (1), with $\mathbb{E}$ replaced by $\mathbb{E}'$. Then, $\sigma^b \in BR\left(\sigma'\right)$ if and only if for $i \in \{1,2\}$ and all $\mathbf{n} \in \mathcal{N}$: $\mathbb{E}'\left(V_i \mid \mathbf{n},s;\sigma'\right) - cn_i\tau > \mathbb{E}'\left(V_{-i} \mid \mathbf{n},s;\sigma'\right) - cn_{-i}\tau \Rightarrow \sigma^b(s,\mathbf{n}) = i$.

Definition 3 (Independent Queues)**.** *A strategy σ^b is a best response strategy for consumers considering queues as independent if, $\sigma^b \in BR\left(\sigma'\right)$ and $f\left(v_1,v_2 \mid \mathbf{n},s;\sigma'\right)$ for any $\mathbf{n}$ reached with probability $q_{i,j}(\mathbf{n},\sigma')\ \forall i,j \in \{1,2\}$ under the strategy σ'.*

Now we apply Definition 3 to calculate consumer strategy based on independent queues.

Proposition 6. *When consumers treat the queues as independent, $\forall \rho > 0, i \in \{1,2\}$ and $n_i - n_{-i} > b_1$, $\sigma^b(s,\mathbf{n}) = i\ \forall s \in \{1,2\}$ for b_1 such that $((\alpha + g(1-\alpha))/(1-g)(1-\alpha))^{b_1-1} = (g/(1-g))(\Delta + b_1 c\tau)/(\Delta - b_1 c\tau)$. If the queue length difference is greater than b_1, the less-informed consumers join the longer queue. Otherwise, they follow their signal.*

Proposition 6 specifies that the queue choice of consumers depends only on the difference between queue lengths. Herd behavior still exists. However, the key difference is that herd behavior persists at all arrival rates. This is because the less-informed consumers assume that all of the other consumers are following their private signals at all states. For any ρ, the queue information is strong since the queue length at any server is interpreted as the number of people with that signal. Thus, for any less-informed consumer, the queue length difference at some threshold, being a collection of additional private signals, is sufficient to overcome any contrary private signal she may possess. Consumers join the longer queue as long as the queue difference between the two queues exceeds a threshold b_1. When the queue length difference is below this threshold, consumers

follow their own signal. (It should be noted that, $\sigma^b(s,\mathbf{n}) \neq \sigma'(s,\mathbf{n})$. Further, the resulting steady state distribution would be inconsistent with the assumed distribution, i.e., $\pi_i(\mathbf{n},\sigma^b) \neq \pi_i(\mathbf{n},\sigma')$). Such consumers (mistakenly) over-attribute higher value to longer queues.

Extent of Learning: As in Section 3.1, we can derive the change in expected valuation for a consumer, who observes queue information (n_1, n_2). For an uninformed consumer ($g = 1/2$), the additional information (from the longer queue) on observing state $\mathbf{n}$ is $V_q(1, s, \mathbf{n}, \sigma^b) = (\Delta/2)\frac{l(\mathbf{n},\sigma^b)-1}{l(\mathbf{n},\sigma^b)+1} = \left(\frac{\Delta}{2}\right)\frac{(1+\alpha)^{n_1-n_2}-(1-\alpha)^{n_1-n_2}}{(1+\alpha)^{n_1-n_2}+(1-\alpha)^{n_1-n_2}}$. For instance, an uninformed consumer arriving at $(11, 10)$ would see an increased valuation of $\alpha\Delta/2$. This is intuitive; since the consumer expects at least α fraction of all consumers to be in the better queue, and she knows that the rest of the consumers are uninformed.

The additional valuation gained by a less-informed consumers with private signal s (with strength $g > 1/2$) by observing a queue i, on observing state $\mathbf{n}$ is $V_q(i, s, \mathbf{n}, \sigma^b)$. It can be shown that (using the likelihood ratios for each state in the expressions in Proposition 5) for the longer queue, the extent of learning is concave increasing in the queue difference $|n_1 - n_2|$. This is consistent with the threshold joining policy observed in equilibrium when consumers treat the queue as being independent.

Given the notion that queue length observation provides additional 'sampling' of private signals, it is not surprising that the 'dagger'-type equilibrium observed for rational consumers disappears. For instance, in Figure 3 (left panel), a rational consumer follows her own signal at $(12, 1)$ but herds at $(11, 0)$, but a consumer under our behavioral consideration would follow identical actions at both states, since her optimal actions are based only on the queue difference and not on the total crowd in the market at the state in which she arrives. Finally, the threshold herding behavior under our behavior consideration continues to exist even when the buffers are asymmetric. The results and related discussions can be found in the Technical Appendix.

4. Regret Minimizing Consumers

In this section, we examine consumers who use decision making criteria other than utility maximization when making their choice, such as minimizing *expected regret* or the *worst-case regret*. Regret can be understood from the following example. For instance, let v_A, v_B be the net value of services from two firms A and B such that $v_A > v_B$ (but unknown to consumers). The consumer chooses firm B, and realizes ex post that the alternative was worse, and regrets not choosing firm A. The regret from choosing firm B is $v_A - v_B$, which equals the expected value that a consumer

would have gained by choosing firm A. There is no regret if firm A is chosen. Thus regret is a measure of *what could have been.*

Consider the decision of any consumer arriving at the market and observing state $\mathbf{n} = (n_1, n_2)$. The consumer has two possible actions at $\mathbf{n} \in \mathcal{N}$ (join queue 1 or queue 2). Suppose the consumer chooses a queue that is revealed to be worse (ex post). The regret this consumer anticipates from a certain choice is the additional expected utility (over all possible realizations) that she would have enjoyed ex post if she had made the alternate (better) choice.

Let j denote the better server, i.e. $j = \arg\max\{V_1, V_2\}$. Consider a consumer who arrives at some state $\mathbf{n} \in \mathcal{N}$ and sees a signal $s \in \{1, 2\}$. Let $R(i|\mathbf{n}, j, s)$ denote the consumer conditional regret. In our context, $R(i|\mathbf{n}, j, s) = \max\{0, (\mathbb{E}[V_{-i}|V_j > V_{-j}] - c\tau(n_{-i} + 1)) - (\mathbb{E}[V_i|V_j > V_{-j}] - c\tau(n_i + 1))\} \ \forall \ i, j, s \in \{1, 2\}$. Then, the expected regret from choosing a queue i at a state $\mathbf{n}$ when seeing a signal s is $ER(i|\mathbf{n}, s) = R(i|\mathbf{n}, 1, s)\Pr(V_1 > V_2|\mathbf{n}, s) + R(i|\mathbf{n}, 2, s)\Pr(V_1 < V_2|\mathbf{n}, s)$. The maximum regret from choosing a queue i at a state $\mathbf{n}$ when seeing a signal s is $MR(i|\mathbf{n}, s) = \max_{j\in\{1,2\}} R(i|\mathbf{n}, j, s)$.

Let $a^{mr}(s, \mathbf{n}), a^{er}(s, \mathbf{n}) \in \{1, 2\}$ be the actions that minimize the maximum regret and expected regret respectively for a consumer $\forall n \in \mathcal{N}, s \in \{1, 2\}$. Then,

$$a^{mr}(s, \mathbf{n}) = \arg\min_{i\in\{1,2\}}\{MR(i|\mathbf{n}, s)\} \quad \text{and} \quad a^{er}(s, \mathbf{n}) = \arg\min_{i\in\{1,2\}}\{ER(i|\mathbf{n}, s)\}.$$

In the following definition, we characterize the regret minimizing queue selection behavior.

Definition 4. (a. Minimize Expected Regret). *The strategy $\sigma^{er}(s, \mathbf{n})$ is an expected regret-minimizing pure strategy equilibrium if $\sigma^{er}(s, \mathbf{n}) \in \arg\min_{i\in\{1,2\}}\{ER(i|\mathbf{n}, s)\}$ and $f\,(v_1, v_2 \mid \mathbf{n}, s; \sigma^{er})$ is defined for any $\mathbf{n}$ that is reached on the equilibrium path with a positive probability.*

(b. Minimax Regret). *A strategy $\sigma^{mr}(s, \mathbf{n})$ minimizes the worst case regret, if $\sigma^{mr}(s, \mathbf{n}) \in \arg\min_{i\in\{1,2\}}\{MR(i|\mathbf{n}, s)\}$ for all $\mathbf{n} \in \mathcal{N}$.*

Consumers who are perfectly informed always choose the server that provides the highest value net of waiting costs, and have no ex post regret. However, the less-informed consumers need to infer quality information from the queue lengths to decide on their regret minimizing strategy. Proposition 7 captures the behavior of the less-informed consumers.

Proposition 7. *(i) For $n \in \mathcal{N}$ and $s \in \{1, 2\}$, $\sigma^{er}(s, \mathbf{n}) = \sigma^*(s, \mathbf{n})$. The queue joining strategies of expected regret minimizing consumers and rational (utility-maximizing) consumers are identical.*

(ii) Consumers who minimize the worst case regret always join the shorter queue, $\sigma^{mr}(s, \mathbf{n}) = i$ if $n_i < n_{-i}$ for $i = 1, 2$. When the queue lengths are equal, consumers are indifferent between the queues.

Proposition 7(i) shows that herd behavior persists under the expected regret minimizing criterion: this finding under a different decision-making approach lends robustness to herd behavior we found under rational Bayesian decision-making approach. Herd behavior continues to persist at low traffic intensity, when consumers minimize expected regret. It is interesting to note that herd behavior is diminished only when consumers minimize the worst case regret. While herding may often lead a consumer to a server with higher value, it may also lead her to the worst disutility, when a consumer joins the longest queue that ex post turns out to have the worst service value. The consumer would then regret not having chosen a better server that was also less congested. Thus, an ex ante examination of the worst case outcomes, leads to a tempered herd behavior. In particular, the minimax regret criterion eliminates herd behavior. In addition, this behavior is independent of the steady state probability distributions.

Modeling consumers as agents who minimize their maximum regret allows us to model consumer decisions when the cognitive costs of utility maximizing behavior are high. Furthermore, modeling minimax *regret* decouples the decision making process from the evaluation of steady state distribution: at every state, regardless of the probability of reaching that state, the consumer evaluates her action based on the regret she anticipates from choosing that action and rejecting the alternatives. With this additional tractability, we find that herd behavior persists in markets with partially unobservable queues (For the sake of brevity, the discussion on the persistence of herd behavior in partially observable queues is deferred to Technical Appendix E.). Finally, Proposition 7(ii) applies as well to convex waiting costs since the shorter queue is more appealing under increased costs (i.e., the threshold for herd behavior increases under convex waiting costs).

5. Conclusions

Many service performance models assume that consumers make decisions about which service to select in a 'vacuum'. In reality, a consumer's decision is influenced by what they observe in other decision-makers around them, and those decisions are manifested through queues. In this paper, we built a model of the choice behavior of consumers when facing congested queues and informational uncertainty. We examined different decision making perspectives on herd behavior in queues under

rationality and regret.

Our paper integrates negative waiting cost externalities due to queueing and positive externalities due to herding behavior, thus bridging both queueing and herding theories in service choice. We use three cognitive models to understand the queue choice of consumers in congestion-prone environments with quality uncertainty, when consumers are (i) rational Bayesian agents, (ii) agents who ignore queue length dependencies, and (iii) *regret* minimizing agents.

We established that the rate at which the consumers arrive to the market greatly influences the service they choose. Holding the arrival rates equal in two markets, herding occurs in a market where service rates are *fast* rather than in a market where service rates are *slow*. With fast service, long queues are more salient than with slow service. In fact, we find that high traffic intensity with slow service reduces herding behavior in finite buffers. In general, queue joining behavior may be complex. In the case of rational Bayesian or expected regret minimizing consumers, a typical queue joining pattern that emerges is the following: consumers use both the size of the crowd in the market relative to the buffer size and the relative allocation of the crowd between the queues in making their queue selection. For instance, rational Bayesian consumers may 'buck the trend' when they see a small crowd, because they impute that the minority waiting in the shorter queue are more informed than the rest. However, if consumers treat the queues as independent queues, they herd according to a threshold policy that depends only on the difference in queue lengths. When consumers minimize the worst case regret, there is no herding and consumers always join the shorter queue.

Our model makes some predictive hypotheses related to how consumers make their choices. An exploration (empirical or experimental) of queue joining behavior along the lines of Schweitzer and Cachon (2000) could reveal how consumers make such tradeoffs in real life. While the research on bounded rationality is rich (for instance, see Su (2008) for bounded rationality in newsvendor models), there is a paucity of theoretical and empirical research on boundedly rational decision making in queues. A way to test for bounded rationality would be to check if consumers, in real settings, ignore the size of the crowd when making their decision about *which* queue to join.

When consumers choose a queue in order to minimize their maximum regret, we note that they join the shorter queue at all states. Thus we find that the classical approach to minimize the expected waiting times (ignoring the service value) is consistent with minimax regret action (when service values are unknown). One hypothesis would be that consumers join shorter queues because they minimize their worst case regret, since the information on service value is usually limited.

When a consumer considers buying a product or service of whose quality she is completely uninformed, she will often herd and wait to consume the service after others. This is consistent with rational utility maximization. On the other hand, regret minimizing consumers may choose the less congested service. Thus, studying consumer choices in congested environments would help firms and researchers identify the nature of the decision models that consumers use to arrive at their decisions.

Appendix A: Computational Issues

We determine the equilibrium strategy profile through the iterative process described in the Table below.

Step 0:	Construct an initial strategy profile σ=σ_0 based on the input parameters.
Step 1:	For the given strategy profile σ, calculate the steady state probabilities $\pi_1(\boldsymbol{n}, \sigma)$ and $\pi_2(\boldsymbol{n}, \sigma)$.[a]
Step 2:	Once the stationary probabilities for profile σ are available, a new profile σ_{new} is constructed by using rational strategy profile derived from the result of Lemma 1.
Step 3:	Check convergence criteria.[b] If satisfied, stop. The equilibrium strategy profile is σ_{new}. Else repeat Step 1 with $\sigma = \sigma_{new}$.

[a] We utilize the regenerative theory based approach (due to Grassmann et al. (1985)) to calculate the stationary vector of the two dimensional birth and death process where simple schemes such as Gaussian elimination fail to converge well. The likelihood ratios $l(\mathbf{n}, \sigma)$ are then calculated using the derived steady state probabilities.

[b] The convergence criteria employed was that the maximum difference between corresponding steady state probabilities at any state between successive iterations was less than 10^{-5}.

Figure A1: *Procedure for the determination of an equilibrium strategy profile for a given g, α, and N.*

Our iterative procedure in Table A1 (for the rational Bayesian consumer model) concludes in several minutes on a PC with 2.4Ghz Intel processor, under our convergence criteria, for different parameter settings, and for different choice of initial strategies. In many cases, the exact equilibrium strategy is found. However, we believe, in worst cases, for *exact* convergence, the procedure may take exponential time. Recent literature demonstrates that the problem of finding just one Nash equilibrium of a finite normal-form game is PPAD-complete (Papadimitriou 2008). PPAD stands for *Polynomial Parity Argument (Directed case)*. For every normal form game with countable strategies Nash equilibrium is guaranteed to exist, whereas in many typical NP-complete problems, the solution sought may or may not exist. To address this difference, an appropriate class of complexity needs to be considered. The complexity class PPAD relates Nash to a host of computational problems with guaranteed existence of solutions. Many problems belonging to the PPAD class are known to be intractable. See Papadimitriou (2008) for a formal definition of PPAD-completeness, and examples of problem instances that belong to this class.

Establishing the computational complexity of finding Nash Equilibria for various games remains

one of the "most important concrete open questions" in theoretical computer science (Papadimitriou 2008). In fact, the standard method for finding Nash equilibrium in bimatrix games called the Lemke-Howson Algorithm (Lemke and Howson 1964) has recently been shown to take exponential number of steps in some cases (Savani and von Stengel 2001). Our simple procedure is also specific to the problem at hand. Still, converging to the exact equilibrium strategy may take an exponential number of iterations in some cases. Since our main focus is on understanding properties of consumer choice behavior in queue settings, establishing computational complexity is beyond the scope of this paper. It may indeed be that our game could be solved efficiently computationally.

Appendix B: Market Share Impact of Herd Behavior

In this appendix, we study the impact of herd behavior on the market share of the better server. W.l.o.g, let server 1 be better than server 2. Consider a market with α fraction of *informed* consumers and $(1-\alpha)$ fraction of uninformed consumers (with signal strength $g = 1/2$). When the consumers select a server based on their private information only, the arrival rates at servers 1 and 2 are $\lambda(\alpha + (1-\alpha)(1/2))$ and $\lambda(1-\alpha)(1/2)$ respectively. We consider the market in Section 3.1. Consider the following markets:

1. M_a: The uninformed consumers herd in this market.
2. M_b: All consumers follow their signals.
3. Market where all consumers follow their signal (but there is no blocking from either queue).

For instance, if all consumers chose the server according to their private signal and no one is blocked, the market share of the better firm would be m. Let $m = \alpha + (1-\alpha)(1/2)$ measures the total "private information" in the market. Let $\lambda_1^*(M_i)$ and $\lambda_2^*(M_i)$ be the equilibrium arrival rates to the better server (server 1) and the inferior server (2) in a market i. We write down the equilibrium market shares of the better server, and plot them in Figure B1.

$$MarketShare(server1) \quad = \quad (\lambda_1^*(M_i))/(\lambda_1^*(M_i) + \lambda_2^*(M_i)) \;\; \forall i \in \{a, b\}$$

Note that the market share of the better service facility in a market with queueing delays can *exceed* its market share achieved in a market without any queueing externalities m. This shows that the market share of a service facility improves in a market with congestion externalities. Also, note that market share improvement for the better server occurs at low arrival rates. The managerial implications suggest that (a) the better server should have much higher capacity than rate of arrivals (to trigger herd behavior), and (b) the better server has an incentive to let differentially-informed consumers communicate their private information (imperfectly) through their actions. What information should a firm share to improve revenues is a non-trivial question. This is because the a high quality firm will have different incentives from a low quality firm in a market where consumers are imperfectly informed. Both firms can influence the information in the market by signaling (by adjusting service rates, for instance). Such a signaling game is considered in Debo and Veeraraghavan (2010).

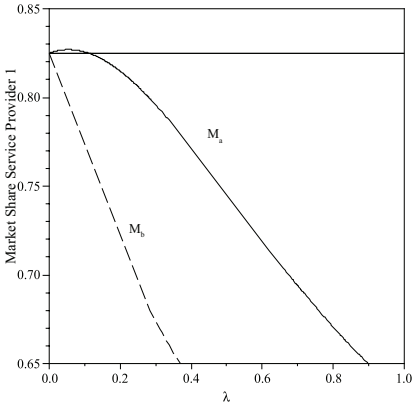

Figure B1: *Market share as a function of the arrival rate in different markets. The parameters are* $m = \alpha + \frac{1}{2}(1-\alpha) = 0.825$ *and* $c = 0.1$*. The horizontal line represents the high quality firm market share without any blocking effects. The dotted curve shows the market share of the firm when all customers follow their signal, with blocking effects (Market* M_b*). The thick curve shows the market share under herd behavior with blocking (Market* M_a*). The figure shows that the better firm gains significant market share in equilibrium (due to herding), compared to the case when there is no herding.*

Appendix C: Proofs

Proof. of **Lemma 1:** For the sake of notational convenience, let us define $\phi = \left(\frac{g}{1-g}\right)$. Furthermore, from the definition of V_+ and V_-, we have that:

$$\mathbb{E}(V_1 \mid \mathbf{n}, 1) = \frac{\phi l(\mathbf{n}) V_+ + V_-}{\phi l(\mathbf{n}) + 1} \text{ and } \mathbb{E}(V_2 \mid \mathbf{n}, 1) = \frac{\phi l(\mathbf{n}) V_- + V_+}{\phi l(\mathbf{n}) + 1}.$$

Thus, the agent joins queue 1 on signal 1 (on signal 2) if:

$$\frac{\phi l(\mathbf{n}) V_+ + V_-}{\phi l(\mathbf{n}) + 1} - \frac{\phi l(\mathbf{n}) V_- + V_+}{\phi l(\mathbf{n}) + 1} > (n_1 - n_2) c\tau \Leftrightarrow \phi l(n_1, n_2) > \frac{\Delta + (n_1 - n_2) c\tau}{\Delta - (n_1 - n_2) c\tau} \Rightarrow \sigma(1, \mathbf{n}) = 1$$
$$\frac{V_- + \frac{l(\mathbf{n})}{\phi} V_+}{1 + \frac{l(\mathbf{n})}{\phi}} - \frac{V_+ + \frac{l(\mathbf{n})}{\phi} V_-}{1 + \frac{l(\mathbf{n})}{\phi}} > (n_1 - n_2) c\tau \Leftrightarrow \frac{l(n_1, n_2)}{\phi} > \frac{\Delta + (n_1 - n_2) c\tau}{\Delta - (n_1 - n_2) c\tau} \Rightarrow \sigma(2, \mathbf{n}) = 1.$$

If following signal 1 is rational in state (n_1, n_2) then following signal 2 is rational in state (n_2, n_1):

$$\frac{1}{\phi l(n_1, n_2)} < \frac{\Delta + (n_2 - n_1) c\tau}{\Delta - (n_2 - n_1) c\tau} \Rightarrow \frac{l(n_2, n_1)}{\phi} < \frac{\Delta + (n_2 - n_1) c\tau}{\Delta - (n_2 - n_1) c\tau},$$

from which follows that following signal 2 is rational in state (n_2, n_1). As the service rates are equal, we have that: $l(0,0) = l(1,1) = 1$. As $(\phi V_+ + V_-)/(\phi + 1) - (V_+ + \phi V_-)/(\phi + 1) = \Delta > 0$, it follows that $\sigma(s, (0,0)) = \sigma(s, (1,1)) = s$, i.e. agents follow their signal when the queue lengths

are equal, provided that positive utility can be obtained from joining $\frac{\phi V_+ + V_-}{\phi+1} > c\tau$. Therefore,

$$\begin{cases} \sigma(s,\mathbf{n})=1 & \Leftrightarrow & \frac{g}{1-g}\frac{\Delta+(n_1-n_2)c\tau}{\Delta-(n_1-n_2)c\tau} & < & l(\mathbf{n};\sigma) & & & (L) \\ \sigma(s,\mathbf{n})=s & \Leftrightarrow & \frac{1-g}{g}\frac{\Delta+(n_1-n_2)c\tau}{\Delta-(n_1-n_2)c\tau} & < & l(\mathbf{n};\sigma) & < & \frac{g}{1-g}\frac{\Delta+(n_1-n_2)c\tau}{\Delta-(n_1-n_2)c\tau} & (F) \\ \sigma(s,\mathbf{n})=2 & \Leftrightarrow & & & l(\mathbf{n};\sigma) & < & \frac{1-g}{g}\frac{\Delta+(n_1-n_2)c\tau}{\Delta-(n_1-n_2)c\tau} & (S)\,. \end{cases}$$

□

Proof. of **Corollary 2:** Symmetric strategies of consumers implies $\frac{1-g}{g} \leq l(n,n) = 1$. Applying Lemma 1, less-informed consumers follow their signal. □

Lemma C1. *All consumers join the shorter queue when $\Delta < c\tau$.*

Lemma C2. *When $\Delta > c\tau$: the likelihood ratios l^L and l^S at $(1,0)$ have the following properties:*

1. $\lim_{\rho\to 0} l^L = \lim_{\rho\to 0} l^S = \frac{1+\alpha}{1-\alpha} > 1$
2. *l^S is decreasing in ρ for all α and* $\lim_{\rho\to\infty} l^S = 1$.
3. *l^L is unimodal in ρ over $\rho \in [0,\infty)$ for any α and achieves the maximum $l^L_{\max} = \max_{\rho>0} l^L$. Further,* $\lim_{\rho\to\infty} l^L = 1$. *Clearly $l^L_{\max} \geq \frac{1+\alpha}{1-\alpha}$.*
4. *The likelihood ratios can be ranked as follows: $l^L \geq l^S \geq 1$ for all ρ, α.*

Proof. of **Lemmas C1** and **C2:** Deferred to the Technical appendix. □

Proof. of **Proposition 3:**

Proposition 3(1.):

(i) When $\alpha\Delta < c\tau < \Delta$. Since, $\alpha\Delta < c\tau$, we have $\alpha < \frac{c\tau}{\Delta}$. Therefore, $\frac{1+\alpha}{1-\alpha} < \frac{\Delta+c\tau}{\Delta-c\tau}$. From item (2) in Lemma C2 we have l^S decreasing in ρ for all $\alpha > 0$. Also, from item 1, $lim_{\rho\to 0} l^S(\alpha,\rho) = \frac{1+\alpha}{1-\alpha}$. Hence, we have $l^S(\alpha,\rho) < \left(\frac{1+\alpha}{1-\alpha}\right)$ for all $\rho > 0$. Since, $\frac{1+\alpha}{1-\alpha} < \frac{\Delta+c\tau}{\Delta-c\tau}$, we have $l^S(\alpha,\rho) < \frac{\Delta+c\tau}{\Delta-c\tau}\ \forall\ \rho > 0$. Therefore, $\sigma^* = \sigma_S\ \forall\ \rho > 0$.

(ii) We have $c\tau < \alpha\Delta$, which implies that $\frac{\Delta+c\tau}{\Delta-c\tau} < \frac{1+\alpha}{1-\alpha}$. Since $lim_{\rho\to 0} l^S(\alpha,\rho) = \frac{1+\alpha}{1-\alpha}$ and we have l^S decreasing in ρ for all $\alpha > 0$, there exists a unique $\overline{\rho}$ such that $l^S(\alpha,\overline{\rho}) = \frac{\Delta+c\tau}{\Delta-c\tau}$. Then, $l^S(\alpha,\rho) \leq \frac{\Delta+c\tau}{\Delta-c\tau}$ for all $\rho \geq \hat{\rho}$. Therefore, for all $\rho \in (\hat{\rho},\infty)$, $l^S(\alpha,\rho) < \frac{\Delta+c\tau}{\Delta-c\tau}$, and from Lemma 1, σ_S is in equilibrium for $\rho \in (\hat{\rho},\infty)$.

Proposition 3(2.):

(i) Proving $\forall\rho \in (\underline{\rho},\overline{\rho})$ when $\alpha\Delta < c\tau < \hat{\alpha}\Delta$ (where $\alpha < \hat{\alpha}$). Given, $\alpha\Delta < c\tau$ we have $\frac{\Delta+c\tau}{\Delta-c\tau} > \frac{1+\alpha}{1-\alpha}$. From Lemma C2 item 3, recall that $l^L(\alpha,\rho)$ is unimodal and reaches a maximum at some ρ_{max}. $l^L(\alpha,\rho)$ is continuous and increasing in $[0,\rho_{max}(\alpha)]$, and $l^L(\alpha,\rho)$ is continuous and decreasing in $[\rho_{max}(\alpha),\infty)$. Also, $lim_{\rho\to 0} l^L(\alpha,\rho) = \frac{1+\alpha}{1-\alpha} > 1$ and $lim_{\rho\to\infty} l^L(\alpha,\rho) = 1$.

Therefore applying Rolle's theorem, there are two points $\underline{\rho}$, $\overline{\rho}$ such that $l^L(\alpha,\underline{\rho}) = l^L(\alpha,\overline{\rho}) = \left(\frac{\Delta+c\tau}{\Delta-c\tau}\right) (> \frac{1+\alpha}{1-\alpha})$. Therefore, $l^L(\alpha,\rho) > \left(\frac{\Delta+c\tau}{\Delta-c\tau}\right)$ for $\rho \in [\underline{\rho},\overline{\rho}]$. Finally, for $\forall\rho$, $l^L(\alpha,\rho_{max}) > l^L(\alpha,\rho)$. Hence, for $\rho \in [\underline{\rho},\overline{\rho}]$, we have, $l^L(\alpha,\rho_{max}) > \left(\frac{\Delta+c\tau}{\Delta-c\tau}\right)$ which implies, $c\tau < \frac{l^L_{\max}-1}{l^L_{\max}+1} \triangleq \hat{\alpha}$.

Hence, for $l^L(\alpha,\rho) \geq \left(\frac{\Delta+c\tau}{\Delta-c\tau}\right)$ for $\rho \in [\underline{\rho},\overline{\rho}]$ when $\alpha\Delta < c\tau < \hat{\alpha}\Delta$. Applying the result of Lemma 1, we have $\sigma^* = \sigma^L$ for $\rho \in (\underline{\rho},\overline{\rho})$.

(ii) $\forall \rho \in (0,\overline{\rho}]$ when $c\tau \leq \alpha\Delta$. When $c\tau \leq \alpha\Delta$, we have $l^L(\alpha,\rho)$ is continuous and increasing in the interval $[0,\rho_{max}(\alpha)]$, and $l^L(\alpha,\rho)$ is continuous and decreasing in the interval $[\rho_{max}(\alpha),\infty)$. Also, $\lim_{\rho\to 0} l^L(\alpha,\rho) = \frac{1+\alpha}{1-\alpha} > 1$ and $\lim_{\rho\to\infty} l^L(\alpha,\rho) = 1$. Hence, $l^L(\alpha,\rho) > \frac{1+\alpha}{1-\alpha}$ for all $\rho \in (0,\overline{\rho})$. When $c\tau \leq \alpha\Delta$, we have $< \frac{1+\alpha}{1-\alpha} \geq \left(\frac{\Delta+c\tau}{\Delta-c\tau}\right)$. This gives, $l^L(\alpha,\rho) > \left(\frac{\Delta+c\tau}{\Delta-c\tau}\right)$ for all $\rho \in (0,\overline{\rho})$. Hence, $\sigma^* = \sigma^L$ for $\rho \in (0,\overline{\rho})$. □

Proof. **of Proposition 4:** Without loss of generality, let $n_1 > n_2$. The expected valuation from queues based on private signals, (before observing the queue lengths) for uninformed consumer is:

$$\mathbb{E}\left(V_1|1\right) = (V_+ + V_-)/2 \text{ and } \mathbb{E}\left(V_2|1\right) = (V_+ + V_-)/2.$$

On observation of state the state $\mathbf{n} = (1,0)$, and based on strategy σ, the updated values for the queues is as follows

$$\mathbb{E}\left(V_1|1;\sigma\right) = \frac{V_+ l\left(\mathbf{n},\sigma\right) + V_-}{l\left(\mathbf{n},\sigma\right)+1} \text{ and } \mathbb{E}\left(V_2|1;\sigma\right) = \frac{V_- l\left(\mathbf{n},\sigma\right) + V_+}{l\left(\mathbf{n},\sigma\right)+1}.$$

The additional valuation for an uninformed consumer at the longer queue (at state (1,0)) is

$$V_q(1,s,(1,0),\sigma) = \mathbb{E}\left(V_1|1;\sigma\right) - \mathbb{E}\left(V_1|1\right) = \left(\frac{V_+ l\left(\mathbf{n},\sigma\right) + V_-}{l\left(\mathbf{n},\sigma\right)+1}\right) - \frac{V_+ + V_-}{2} = \frac{l\left((1,0),\sigma\right)-1}{l\left((1,0),\sigma\right)+1}(\Delta/2)\ \forall s \in \{1,2\}.$$

Solving for steady state probabilities, we get

$$V_q(1,s,\mathbf{n},\sigma^{\mathbf{L}}) = \frac{\alpha\Delta}{2}\left(\frac{16.+36\rho+30\rho^2+2\rho^2\alpha+11\rho^3}{16+8\rho\alpha+28\rho+10\rho^2\alpha+28\rho^2+2\rho^2\alpha^2+3\rho^3\alpha+20\rho^3+6\rho^4}\right)\forall s \in \{1,2\} \text{ and}$$

$$V_q(1,s,\mathbf{n},\sigma^{\mathbf{S}}) = \frac{\alpha\Delta}{2}\left(\frac{16+36\rho+34\rho^2+11\rho^3-2\rho^2\alpha}{16+44\rho-8\rho\alpha-10\rho^2\alpha+48\rho^2-3\rho^3\alpha+26\rho^3+6\rho^4+2\rho^2\alpha^2}\right)\ \forall s \in \{1,2\}$$

Similar to the analysis in the proof of Lemma C2, we can show that for $s \in \{1,2\}$,

1. $\lim_{\rho\to 0} V_q(1,s,\mathbf{n},\sigma^{\mathbf{L}}) = \lim_{\rho\to 0} V_q(1,s,\mathbf{n},\sigma^{\mathbf{S}}) = \frac{\alpha\Delta}{2} > 0$.
2. $V_q(1,s,\mathbf{n},\sigma^{\mathbf{S}})$ is decreasing in $\rho \geq 0$ for all α and $\lim_{\rho\to\infty} V_q(1,s,\mathbf{n},\sigma^{\mathbf{S}}) = 0$.
3. $V_q(1,s,\mathbf{n},\sigma^{\mathbf{L}})$ is unimodal in $\rho \in [0,\infty)$ for any α and achieves the maximum at some $\rho = \rho_{\max}$. Further, $\lim_{\rho\to\infty} V_q(1,s,\mathbf{n},\sigma^{\mathbf{L}}) = 0$. Clearly at $\rho = \rho_{\max}$, $V_q(1,s,\mathbf{n},\sigma^{\mathbf{L}}) \geq \frac{\alpha\Delta}{2}$.
4. $V_q(1,s,\mathbf{n},\sigma^{\mathbf{L}}) \geq V_q(1,s,\mathbf{n},\sigma^{\mathbf{S}})$ for all ρ,α.

Using items 1 and 2, we know that $\max_{\rho\geq 0}\{V_q(1,s,\mathbf{n},\sigma^{\mathbf{S}})\} = (\alpha\Delta/2)$. Then using item 3 it follows that there exists a $\overline{\rho}$ such that $V_q(1,s,\mathbf{n},\sigma^{\mathbf{L}}) \geq (\alpha\Delta/2)\ \ \forall \rho \in [0,\overline{\rho}]$. □

Proof. **of Proposition 5:** Let l^{XY} represent the likelihood ratio at state **1** where X, Y ($X, Y \in \{L, S, F\}$) where represents the strategies of consumer class $1, 2$ respectively. It is straightforward to show that the informed consumers always follow their signal since $l^{FY} < \infty (= \frac{1}{1-1}\frac{\Delta+c\tau}{\Delta-c\tau})$ for any less-informed consumer strategy Y. For simplicity, let us represent the likelihood ratios at with strategies Y by l^Y. (i.e. simply, l^{FF} becomes l^F). Let us consider the likelihood ratio under the strategy (F, L). $l^L = N(g, \alpha, \rho)/D(g, \alpha, \rho)$ where

$$\begin{aligned} N(g,\alpha,\rho) &= -16g - 16\alpha + 16\alpha g - 28g\rho - 36\alpha\lambda + 28g\lambda\alpha - 20\rho^2 g + 20\rho^2\alpha g - 4\alpha\rho^2 g + \\ &\quad 4\alpha^2\rho^2 g - 4\rho^2 - 28\alpha\rho^2 - 4\alpha^2\rho^2 - 8\rho^3 g - 6\rho^3 - 11\alpha\rho^3 + 8\rho^3\alpha g - 3\rho^4 \text{ and} \\ D(g,\alpha,\rho) &= -28\rho - 16 - 3\rho^4 - 14\rho^3 - 16\rho^2\alpha g + 11\alpha\rho^3 + 32\alpha\rho^2 + 36\alpha\rho \\ &\quad -8\rho\alpha - 24\rho^2 + 16g - 3\rho^3\alpha - 12\rho^2\alpha - 4\alpha^2\rho^2 g - 4\alpha\rho^2 g + 4\alpha^2\rho^2 g - 4 \\ &\quad \alpha^2\rho^2 + 4\rho^2\alpha^2 + 16\alpha - 16\alpha g + 28g\rho + 20\rho^2 g + 8\rho^3 g - 8\rho^3\alpha g - 28g\rho\alpha. \end{aligned}$$

For the less-informed consumers to join the longer queue, we require $l^L > \frac{g}{1-g}\frac{\Delta+c\tau}{\Delta-c\tau}$. Since l^L is decreasing and $\left(\frac{\alpha+(1-\alpha)g}{1-\alpha-(1-\alpha)g}\right) > \frac{g}{1-g}$, we have $l^L > \frac{g}{1-g}\frac{\Delta+c\tau}{\Delta-c\tau}$ for all $\rho \in [0, \rho^*)$ for some ρ^* such that $l^L(g, \alpha, \rho^*) = \frac{\Delta+c\tau}{\Delta-c\tau}\frac{g}{1-g}$. (By continuity, $\exists \eta^*$ such that $\rho \leq \rho^*$ for $\frac{\Delta+c\tau}{\Delta-c\tau} > \eta^*$. Therefore, the less-informed consumers do not join the longer-queue for any $\frac{\Delta+c\tau}{\Delta-c\tau} > \eta^*$, i.e., for any $c > c^*$). Hence the less-informed consumers join the longer queue, when waiting costs are lower than some c^* and for all ρ less than ρ^*.
For the second part,

$$\begin{aligned} V_q(1, s, \mathbf{n}, \sigma) &= \mathbb{E}\left(V_1 | s; \sigma\right) - \mathbb{E}\left(V_1 | s\right), \\ V_q(1, 1, \mathbf{n}, \sigma^L) &= \frac{gl^L V_+(1-g)V_-}{gl^L + (1-g)} - \frac{gV_+ + (1-g)V_-}{gl^L + (1-g)} = g(1-g)\Delta\frac{(l^L - 1)}{(gl^L + (1-g))}, \\ V_q(1, 2, \mathbf{n}, \sigma^L) &= \frac{gV_+(1-g)l^L V_+}{g + l^L(1-g)} - \frac{gV_- + (1-g)V_+}{gl^L + (1-g)} = g(1-g)\Delta\frac{(l^L - 1)}{(g + l^L(1-g))}. \end{aligned}$$

Similarly, for the additional valuation learned from the shorter queue $(V_q(2, s, \mathbf{n}, \sigma)),\ s \in \{1, 2\}$,

$$V_q(2, 1, \mathbf{n}, \sigma^L) = g(1-g)\Delta\frac{(1 - l^L)}{(gl^L + (1-g))} \text{ and } V_q(2, 2, \mathbf{n}, \sigma^L) = g(1-g)\Delta\frac{(l^L - 1)}{(g + l^L(1-g))}. \quad (5)$$

Using an approach similar to Lemma C2 and the above steps, we can show that $l^L > l^F > l^S$ for $\rho < \overline{\rho}$ which concludes the proof. □

Proof. of **Proposition 6:** The consumers impose the following simplification in their beliefs: The fully informed consumers choose the better queue, and less-informed consumers follow their signal (of strength g). They compute $q_{i,j}(n, \sigma')$ where $\sigma' = \{(\sigma_k) | \sigma_k(s, \mathbf{n}) = s, k \in \{1, 2\}\}$, and $q_{i,j}(n, \sigma')$ is the steady state probability of state n in queue i when $V_j > V_{-j}$. This implies that, under σ', when $V_1 > V_2$, fraction $\alpha + (1-\alpha)g$ consumers join queue 1 and the rest join queue 2. Similarly,

when $V_2 > V_1$, fraction $\alpha+(1-\alpha)g$ join queue 2. Let $g' = \alpha+g(1-\alpha)$. Then $1-g' = (1-g)(1-\alpha)$.

$$\begin{aligned} \text{Then we have, } q_{1,1}(n,\sigma') = q_{1,1}(n,\sigma') &= \frac{(g'\rho)^{n_1}(1-g'\rho)}{(1-(g'\rho)^{N+1})} \text{ and} \\ q_{1,2}(n,\sigma') = q_{2,1}(n,\sigma') &= \frac{((1-g')\rho)^{n_2}(1-(1-g')\rho)}{(1-((1-g')\rho)^{N+1})} \end{aligned}$$

Without loss of generality let $n_1 \geq n_2$. Then for all $(n_1, n_2) \in \mathcal{N}$,

$$\begin{aligned} \pi_1(n_1,n_2) = q_{1,1}(n_1,\sigma') \cdot q_{2,1}(n_2,\sigma') \quad &\text{and} \quad \pi_2(n_1,n_2) = q_{1,2}(n_1,\sigma') \cdot q_{2,2}(n_2,\sigma') \\ l(n_1,n_2) = \frac{\pi_1(n_1,n_2)}{\pi_2(n_1,n_2)} \quad &= \quad (\frac{g'}{1-g'})^{n_1-n_2} = (\frac{\alpha+g(1-\alpha)}{(1-g)(1-\alpha)})^{n_1-n_2}. \end{aligned}$$

Applying the above result to Lemma 1, we find that the queue joining policy is of threshold type. Let $b_1 = (n_1 - n_2)$ be the solution to $(\frac{\alpha+g(1-\alpha)}{(1-g)(1-\alpha)})^(n_1 - n_2) = \frac{g(\Delta+(n_1-n_2)c\tau)}{(1-g)(\Delta-(n_1-n_2)c\tau)}$. It follows that when $n_1 - n_2 \geq b_1$, consumers always join the longer queue. When the queue length difference is less than b_1, they follow their own private signal. When $\alpha = 0$, the above expression for b_1 simplifies to $(\frac{g}{1-g})^{b_1-1} = \frac{\Delta+b_1c\tau}{\Delta-b_1c\tau}$. □

Proof. of **Proposition 7:**
Suppose a less-informed consumer arrives to find the queue lengths to be (n_1, n_2). For sake of simplicity, let high value service be $v_h = \max\{V_1, V_2\}$ and low value service be $v_l = \max\{V_1, V_2\}$. Suppose the consumer chooses queue i (which is revealed to be worse ex post). We denote the regret from a choice i as $R(i)$. Regret is the additional expected utility she would have enjoyed if she had made the alternate (better) choice. Let $R(i|\mathbf{n}, j, s)\ \forall\ \mathbf{n} \in \mathcal{N}, j, s \in \{1, 2\}$ denote the conditional regret a consumer that expects at state $\mathbf{n}$, when her private signal is s and j denotes the better server. In our context, $R(i|\mathbf{n}, j, s) = \max\{0, (\mathbb{E}[V_{-i}|s, V_j > V_{-j}] - c\tau(n_{-i}+1)) - (\mathbb{E}[V_i|s, V_j > V_{-j}] - c\tau(n_i+1))\}\ \forall\ i, \mathbf{n} \in \mathcal{N}, s \in \{1, 2\}, j = \{1, 2\}$.

We now examine possible regret realizations (conditional on the state of the market), when a consumer chooses a queue i: (i) Service i is better and has a shorter queue (no regret). (ii) Service i is better, but the queue was longer. However, the higher value of the service was more than the additional waiting costs (no regret). (iii) Service i is better, but the queue was longer. Thus additional waiting costs negate any benefits from high service value. The regret is $(\mathbb{E}[V_{-i}|V_i > V_{-i}] - c\tau(n_{-i}+1)) - (\mathbb{E}[V_i|V_i > V_{-i}] - c\tau(n_i+1))$. (iv) Service i is worse and had a longer queue. The regret is $(\mathbb{E}[V_{-i}|V_i < V_{-i}] - c\tau(n_{-i}+1)) - (\mathbb{E}[V_i|V_i < V_{-i}] - c\tau(n_i+1))$. (v) Service i is worse but also has a shorter queue. However, the waiting costs savings cannot make up for low value. The regret is $(\mathbb{E}[V_{-i}|V_i < V_{-i}] - c\tau(n_{-i}+1)) - (\mathbb{E}[V_i|V_i < V_{-i}] - c\tau(n_i+1))$. (vi) Service i is worse and has a shorter queue, but the lower waiting costs compensated for the lower service value (no regret).

Note: For sake of analytical convenience, we first prove the result for minimax regret, and then analyze expected regret.

Minimax Regret: The maximum regret in the context of two observable queues for any consumer

is, $MR(i|\mathbf{n},s) = \max_{j\in\{1,2\}}\{R(i|\mathbf{n},j,s)\}$. Let $a^{mr}(s,\mathbf{n}) \in \{1,2\}$ be the action that minimizes the maximum regret $\forall s \in \{1,2\}, \forall k$, i.e.,

$$a^{mr}(s,\mathbf{n}) = \arg\min_{i\in\{1,2\}}\{\max_{j\in\{1,2\}} R(i|\mathbf{n},j,s)\}\forall k.$$

Considering cases (i) through (vi) above, we have

$$\begin{aligned} a^{mr}(s,\mathbf{n}) &= \arg\min_{i\in\{1,2\}}\{\max\{0,0,V_- - c\tau(n_{-i}+1) - (V_+ - c\tau(n_i+1)), \\ &\quad V_+ - c\tau(n_{-i}+1) - (V_- - c\tau(n_i+1)), V_+ - c\tau(n_{-i}+1) - (V_- - c\tau(n_i+1)), 0\}\} \\ &= \arg\min_{i\in\{1,2\}}\{\max\{0, -\Delta - c(n_{-i}-n_i)\tau, \Delta - c(n_{-i}-n_i)\tau\}\} \end{aligned}$$

Without loss of generality, let $n_1 > n_2$. First we examine $MR(1|\mathbf{n},\mathbf{s})$. Note that only cases (ii), (iii) and (iv) apply.

$$MR(1|\mathbf{n},s) = \max\{0, -\Delta - c(n_2-n_1)\tau, \Delta - c(n_2-n_1)\tau\} = (\Delta + c(n_1-n_2)\tau)$$

Similarly, we examine the regret from choosing the shorter queue: $MR(2|\mathbf{n},\mathbf{s})$. Note that only the cases (i), (v) and (vi) apply.

$$MR(2|\mathbf{n},s) = \max\{0, \Delta - c(n_1-n_2)\tau, 0\} = \max\{0, \Delta - c(n_1-n_2)\tau\}$$

Since $(\Delta + c(n_1-n_2)\tau) > \max\{0, \Delta - c(n_1-n_2)\tau\}$, a consumer applies the minimax criterion and will *always* choose queue 2 when $n_1 > n_2$. Thus, when consumers minimizing max regret criterion will always join the shorter queue. They are indifferent between the queues when the two queue lengths are equal. Therefore, minimizing maximum regret leads to the shorter queue joining behavior.

Minimize Expected Regret:

Let $a^{er}(s,\mathbf{n})$ be the action that minimizes expected regret for a consumer arriving at n with signal $s \in \{1,2\}$, i.e., $a^{er}(s,\mathbf{n}) = \arg\min_{i\in\{1,2\}} ER(i|\mathbf{n},s)$

Without loss of generality, let us consider $n_1 > n_2$. First, we note that in case (iii) need not be considered, since the queue is so long as to be insufficient to overcome the additional value, from a better queue. So, consumers will always join the shorter queue when $\Delta - (n_1-n_2)c\tau < 0$. Now, we focus our attention on $\Delta - (n_1-n_2)c\tau$.

$$\begin{aligned} ER(i|n,s) &= R(i|n,1,s)\Pr[V_1 > V_2|n,s] + R(i|n,2,s)\Pr[V_1 < V_2|n,s] \\ &= R(i|n,1,s)\frac{\Pr[n,s|V_1 > V_2]}{\Pr[n,s]} + R(i|n,2,s)\frac{\Pr[n,s|V_1 < V_2]}{\Pr[n,s]}. \end{aligned}$$

$$ER(1|\mathbf{n},1) = \frac{(\Delta + (n_1-n_2)c\tau)(1-g)\pi_2(\mathbf{n})}{g\pi_1(\mathbf{n}) + (1-g)\pi_2(\mathbf{n})} \text{ and } ER(2|\mathbf{n},1) = \frac{(\Delta - (n_1-n_2)c\tau)g\pi_1(\mathbf{n})}{g\pi_1(\mathbf{n}) + (1-g)\pi_2(\mathbf{n})},$$

$$ER(1|\mathbf{n},2)=\frac{(\Delta+(n_1-n_2)c\tau)g\pi_2(\mathbf{n})}{(1-g)\pi_1(\mathbf{n})+g\pi_2(\mathbf{n})} \text{ and } ER(2|\mathbf{n},2)=\frac{(\Delta-(n_1-n_2)c\tau)(1-g)\pi_1(\mathbf{n})}{(1-g)\pi_1(\mathbf{n})+g\pi_2(\mathbf{n})}.$$

For a consumer to join the queue 1 at a state $\mathbf{n}$, regardless of signal s, we need,
$(\Delta+(n_1-n_2)c\tau)(1-g)\pi_2(\mathbf{n})<(\Delta-(n_1-n_2)c\tau)g\pi_1(\mathbf{n})$ and
$(\Delta+(n_1-n_2)c\tau)g\pi_2(\mathbf{n})<(\Delta-(n_1-n_2)c\tau)(1-g)\pi_1(\mathbf{n})$, which gives,

$$l(\mathbf{n})>\left(\frac{g}{1-g}\right)\left(\frac{\Delta+c(n_1-n_2)\tau}{\Delta-c(n_1-n_2)\tau}\right).$$

Thus for given $l(\mathbf{n};\sigma)$, assume that $n_1>n_2$, and let $\sigma^{er}(s,\mathbf{n})=i$ denote that the consumer joins the queue i after observing state $\mathbf{n}$ to minimize regret (i.e. $a^{er}(s,\mathbf{n})=i$).

$$\begin{cases} \sigma^{er}(s,\mathbf{n})=1 & \Leftrightarrow & \frac{g}{1-g}\frac{\Delta+(n_1-n_2)c\tau}{\Delta-(n_1-n_2)c\tau} & < & l(\mathbf{n};\sigma) & & & (L) \\ \sigma^{er}(s,\mathbf{n})=s & \Leftrightarrow & \frac{1-g}{g}\frac{\Delta+(n_1-n_2)c\tau}{\Delta-(n_1-n_2)c\tau} & < & l(\mathbf{n};\sigma) & < & \frac{g}{1-g}\frac{\Delta+(n_1-n_2)c\tau}{\Delta-(n_1-n_2)c\tau} & (F) \\ \sigma^{er}(s,\mathbf{n})=2 & \Leftrightarrow & & & l(\mathbf{n};\sigma) & < & \frac{1-g}{g}\frac{\Delta+(n_1-n_2)c\tau}{\Delta-(n_1-n_2)c\tau} & (S) \end{cases}$$

This condition is identical to the one derived in Lemma 1 for utility maximizing consumers. Finally, the results can be shown to hold for asymmetric buffers in the same fashion. □

References

Banerjee, A. 1992. A Simple Model of Herd Behavior. *Quart. J. of Economics*, 107, pp. 797–818.

Becker, G. 1991. A Note on Restaurant Pricing and Other Examples of Social Influences on Price. *Journal of Political Economy*, 99, (5), pp. 1109-16.

Bell, D. E. 1982. Regret in Decision-Making under Uncertainty. *Operations Research*, Vol 30, 5, pp. 961–981.

Bikhchandani, S., D. Hirshleifer and I. Welch. 1992. A Theory of Fads, Fashion, Custom and Cultural change as Information Cascades. *Journal of Political Economy*, 100, pp. 992–1026.

Callander, S. and J. Horner. 2009. The Wisdom of the Minority. *Journal of Economic Theory*, 144, pp. 1421–39.

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

Debo, L., C. Parlour and U. Rajan. 2007. The Value of Congestion, Chicago Booth Working Paper.

Debo, L., S. Veeraraghavan. 2010. Price and Occupancy Levels as Signals of Quality. Chicago Booth Working Paper.

Fudenberg, D. and J. Tirole. 1991. Game Theory. Cambridge: The M.I.T. Press.

Gilboa, I., and E. Zemel. 1989. Nash and correlated equilibria: Some complexity considerations. *Games and Economic Behavior*, 1, pp. 80–93.

Grassmann, W. K., M.I. Taksar, D.P. Heyman. 1985. Regenerative analysis and steady state distributions for Markov chains, *Operations Research*, Vol 33, pp 1107–1116.

Halfin, S. 1985. The shortest queue problem. *J. Appl. Prob.* 22, pp. 865–878.

Hassin, R and M. Haviv. 2003. To Queue or not to Queue: Equilibrium behavior in queuing systems. Kluwer Academic Publishers, Norwell, MA.

Kingman, J. F. C., 1961. Two similar queues in parallel. *Annals of Math. Stats.*, 32, pp. 1314–1323.

Lemke, C. E. and J. T. Howson, Jr. 1964. The Equilibrium Points of Bimatrix Games. *Journal of the Society for Industrial and Applied Mathematics*, 12, pp. 413–423.

Loomes, G. and R. Sugden, 1982. Regret Theory: An Alternative Theory of Rational Choice Under Uncertainty. *The Economic Journal*, Vol 92, No 368, pp. 805–824.

Luce, R.D. 1959. Individual Choice Behavior: A Theoretical Analysis. Wiley, New York, NY.

Maskin, E. and J. Tirole. 2001. Markov Perfect Equilibrium. Observable action. *Journal of Economic Theory*. 100, 191–219.

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

Papadimitriou, C. P. 2008. The Complexity of finding Nash Equilibria. Algorithmic Game Theory. Cambridge Press. pp 29-50.

Rubinstein, A. 1998. Modeling Bounded Rationality. MIT Press. Cambridge, MA.

Savage, L. 1951. The Theory of Statistical Decision, *Journal of the American Statistical Association*, Vol. 46, No. 253. pp. 55–67.

Savani, R. and B. von Stengel. 2006. Hard-to-solve bimatrix games. *Econometrica*, 74:397–429.

Schweitzer, M. E. and G. P. Cachon. 2000. Decision Bias in the Newsvendor Problem with a Known Demand Distribution: Experimental Evidence. *Management Science*. 46(3), pp. 404–420.

Simon, H. A. 1955. A Behavioral model of Rational Choice. *Quart. J. of Econ.*, 69(1), pp. 99–118.

Su, X. 2008. Bounded Rationality in Newsvendor Models. *Manufacturing & Service Operations Management*, 10(4), pp. 566–589.

Surowiecki, J. 2005. The Wisdom of Crowds. Anchor Publishing. USA

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

Whinston, W. 1977. The Optimality of Joining the Shortest Queue Discipline, *Journal of Applied Probability*, 14, 181–189.

Wolff, R.W. 1982. Poisson Arrivals see Time Averages. *Operations Research*, 30 (2) pp. 223–231.

Online Appendix for
Herding in Queues with Waiting Costs: Rationality and Regret

Lemma T3. *All consumers join the shorter queue when $\Delta < c\tau$.*

Proof. We establish that *all* consumers join the shorter queue when $\Delta < c\tau$. This result helps us focus on the case when $\Delta > c\tau$, when consumers could possibly join the longer queue.

To join the longer queue (server 1), the consumer must have

$$\begin{aligned}
\mathbb{E}(V_1|\mathbf{1};\sigma) - 2c\tau &> \mathbb{E}(V_2|\mathbf{1};\sigma) - c\tau \\
\frac{V_+ l^L(\mathbf{n},\sigma_L) + V_-}{l^L(\mathbf{n},\sigma_L)+1} - 2c\tau &> \frac{V_- l^L(\mathbf{n},\sigma_L) + V_+}{l^L(\mathbf{n},\sigma_L)+1} - c\tau \\
\frac{l^L(\mathbf{n},\sigma_L)-1}{l^L(\mathbf{n},\sigma_L)+1} &> \frac{c\tau}{\Delta} \quad \text{Since } \Delta = V_+ - V_- \\
\text{When } \Delta > c\tau \quad \sigma^* = \sigma_L &\Leftrightarrow l^L(\mathbf{n},\sigma_L) > \frac{\Delta + c\tau}{\Delta - c\tau} \qquad (6) \\
\Delta < c\tau \quad \sigma^* = \sigma_L &\Leftrightarrow l^L(\mathbf{n},\sigma_L) < \frac{\Delta + c\tau}{\Delta - c\tau} \qquad (7)
\end{aligned}$$

Since $l^L\left(\mathbf{n},\sigma^L\right) > 0$ and $\frac{\Delta+c\tau}{\Delta-c\tau} < 0$ when $\Delta < c\tau$, therefore $\sigma^* \neq \sigma^L$ when $\Delta < c\tau$.

Similarly, the consumer joins the shorter queue when,

$$\begin{aligned}
\mathbb{E}(V_1|\mathbf{1};\sigma) - 2c\tau &< \mathbb{E}(V_2|\mathbf{1};\sigma) - c\tau \\
\frac{l^S(\mathbf{n},\sigma_S)-1}{l^S(\mathbf{n},\sigma_S)+1} &< \frac{c\tau}{\Delta} \\
\text{when } \Delta > c\tau \qquad \sigma^* = \sigma_S &\Leftrightarrow l^S(\mathbf{n},\sigma_S) < \frac{\Delta + c\tau}{\Delta - c\tau} \\
\Delta < c\tau \qquad \sigma^* = \sigma_S &\Leftrightarrow l^S(\mathbf{n},\sigma_S) > \frac{\Delta + c\tau}{\Delta - c\tau} > 0 \qquad (8)
\end{aligned}$$

The consumer always joins the shorter queue when $\Delta < c\tau$. □

Lemma T4. *When $\Delta > c\tau$: the likelihood ratios l^L and l^S at $(1,0)$ have the following properties:*

1. $\lim_{\rho\to 0} l^L = \lim_{\rho\to 0} l^S = \frac{1+\alpha}{1-\alpha} > 1$
2. *l^S is decreasing in ρ for all α and* $\lim_{\rho\to\infty} l^S = 1$.
3. *l^L is unimodal in ρ over $\rho \in [0,\infty)$ for any α and achieves the maximum $l^L_{\max} = \max_{\rho>0} l^L$. Further,* $\lim_{\rho\to\infty} l^L = 1$. *Clearly $l^L_{\max} \geq \frac{1+\alpha}{1-\alpha}$.*
4. *The likelihood ratios can be ranked as follows: $l^L \geq l^S \geq 1$ for all ρ, α.*

Proof. First, let $\Delta > c\tau$.

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 $V_i > V_{-i}$. We write out the steady state balance equations for all states when $V_1 > V_2$.

$$
\begin{aligned}
\lambda\pi_1(0,0) &= (1/\tau)(\pi_1(1,0) + \pi_1(0,1)) \\
2(1/\tau)\pi_1(2,2) &= \lambda(\pi_1(1,2) + \pi_1(2,1)) \\
(2(1/\tau) + \lambda)\pi_1(2,1) &= \lambda((1/2)(1-\alpha) + \alpha))\pi_1(1,1) + \lambda\pi_1(2,0) + (1/\tau)\pi_1(2,2) \\
(2(1/\tau) + \lambda)\pi_1(1,2) &= \lambda(1/2)(1-\alpha)\pi_1(1,1) + \lambda\pi_1(0,2) + (1/\tau)\pi_1(2,2) \\
(\lambda + (1/\tau))\pi_1(0,1) &= (1/2)(1-\alpha)\lambda\pi_1(0,0) + (1/\tau)\pi_1(1,1) + (1/\tau)\pi_1(0,2) \\
(\lambda + (1/\tau))\pi_1(1,0) &= ((1/2)(1-\alpha) + \alpha)\lambda\pi_1(0,0) + (1/\tau)\pi_1(1,1) + (1/\tau)\pi_1(2,0) \\
(\lambda + (1/\tau))\pi_1(2,0) &= (\Lambda_1 + \alpha\lambda)\pi_1(1,0) + (1/\tau)\pi_1(2,1) \\
(\lambda + (1/\tau))\pi_1(0,2) &= \Lambda_1'\pi_1(0,1) + (1/\tau)\pi_1(1,2) \\
(\lambda + 2(1/\tau))\pi_1(1,1) &= ((1-\alpha)\lambda - \Lambda_1)\pi_1(1,0) + (\lambda - \Lambda_1')\pi_1(0,1) + (1/\tau)(\pi_1(1,2) + \pi_1(2,1))
\end{aligned}
$$

In the system of equations, we use Λ_1 and Λ_1' to denote the rate of consumers who arrive and choose to join the longer queue at $\mathbf{1} = (1,0)$ and $\mathbf{1}' = (0,1)$, respectively. Therefore, if all uninformed consumers join the longer queue, we have $\Lambda_1 = \Lambda_1' = (1-\alpha)\lambda$ and when all uninformed consumers join the shorter queue, we have $\Lambda_1 = \Lambda_1' = 0$.

A similar set of equations can be written for the case when $V_2 > V_1$. Solving for the steady state probabilities, we obtain the following likelihood ratios, after inserting $\rho = \lambda\tau$.

$$
\begin{aligned}
l^L(\alpha,\rho) = l\left(\mathbf{1};\sigma^{\mathbf{L}}\right) &= \frac{\pi_1(1,0)}{\pi_2(1,0)}(\Lambda_1 = \Lambda_1' = (1-\alpha)\lambda) \\
&= \frac{(3\rho^4 + 10\rho^3 + 7\rho^3\alpha + 8 + 20\rho^2\alpha + 2\alpha^2\rho^2 + 14\rho^2 + 22\alpha\rho + 14\rho + 8\alpha)}{(3\rho^4 + 10\rho^3 - 4\rho^3\alpha + 8 + 14\rho^2 - 10\rho^2\alpha + 14\rho - 14\alpha\rho - 8\alpha)} \\
l^S(\alpha,\rho) = l(\mathbf{1};\sigma^{\mathbf{S}}) &= \frac{\pi_1(1,0)}{\pi_2(1,0)}(\Lambda_1 = \Lambda_1' = 0) \\
&= \frac{3\rho^4 + 13\rho^3 + 4\rho^3\alpha + 8 + 24\rho^2 + 12\rho^2\alpha + 22\rho + 14\alpha\rho + 8\alpha}{(3\rho^4 + 13\rho^3 - 7\rho^3\alpha + 8 + 24\rho^2 + 2\alpha^2\rho^2 - 22\rho^2\alpha - 22\alpha\rho + 22\rho - 8\alpha)}
\end{aligned}
$$

1. $\lim_{\rho\to 0} l^L(\alpha,\rho) = \lim_{\rho\to 0} l^S(\alpha,\rho) = \frac{1+\alpha}{1-\alpha} > 1$. It is easy to note that

$$
\lim_{\rho\to 0} l^L(\alpha,\rho) = \lim_{\rho\to 0} \frac{(3\rho^4 + 10\rho^3 + 7\rho^3\alpha + 8 + 20\rho^2\alpha + 2\alpha^2\rho^2 + 14\rho^2 + 22\alpha\rho + 14\rho + 8\alpha)}{(3\rho^4 + 10\rho^3 - 4\rho^3\alpha + 8 + 14\rho^2 - 10\rho^2\alpha + 14\rho - 14\alpha\rho - 8\alpha)} = \frac{(1+\alpha)}{(1-\alpha)}.
$$

$$
\lim_{\rho\to 0} l^S(\alpha,\rho) = \lim_{\rho\to 0} \frac{3\rho^4 + 13\rho^3 + 4\rho^3\alpha + 8 + 24\rho^2 + 12\rho^2\alpha + 22\rho + 14\alpha\rho + 8\alpha}{(3\rho^4 + 13\rho^3 - 7\rho^3\alpha + 8 + 24\rho^2 + 2\alpha^2\rho^2 - 22\rho^2\alpha - 22\alpha\rho + 22\rho - 8\alpha)} = \frac{(1+\alpha)}{(1-\alpha)}.
$$

2. $l^S(\alpha,\rho)$ is decreasing in ρ for all α.

$$\frac{dl^S(\alpha,\rho)}{d\rho} = \frac{\left[\begin{array}{c} 128\rho\alpha^2 - 33\rho^6\alpha - 204\rho^5\alpha - 502\rho^4\alpha - 64\alpha + 64\alpha^2 - 644\rho^3\alpha + +72\rho^2\alpha^2 \\ -476\rho^2\alpha - 224\rho\alpha + 12\rho^5\alpha^2 + 22\rho^4\alpha^2 + 8\rho^4\alpha^3 + 20\rho^3\alpha^2 - 28\rho^2\alpha^3 - 32\rho\alpha^3 \end{array}\right]}{(3\rho^4 + 13\rho^3 + 8 - 7\rho^3\alpha + 24\rho^2 + 2\rho^2\alpha^2 - 22\rho^2\alpha - 22\rho\alpha + 22\rho - 8\alpha)^2} < 0 \ \forall \alpha \in [0,1].$$

3. $l^L(\alpha,\rho)$ is unimodal in ρ over $[0,\infty)$ and reaches its maximum at some $\rho = \rho_{\max}(\alpha) > 0$.

$$\begin{aligned}\frac{dl^L(\alpha,\rho)}{d\rho} &= \frac{\left[\begin{array}{c} -128\rho\alpha^2 - 33\rho^6\alpha - 180\rho^5\alpha - 470\rho^4\alpha + 64\alpha - 64\alpha^2 - 604\rho^3\alpha \\ -104\rho^2\alpha^2 - 300\rho^2\alpha + 32\rho\alpha - 12\rho^5\alpha^2 - 10\rho^4\alpha^2 + 8\rho^4\alpha^3 - 20\rho^3\alpha^2 - 28\rho^2\alpha^3 - 32\rho\alpha^3 \end{array}\right]}{(3\rho^4 + 10\rho^3 - 4\rho^3\alpha + 8.0 + 14\rho^2 - 10\rho^2\alpha + 14\rho - 14\rho\alpha - 8\alpha)^2} \\ &= \frac{G(\alpha,\rho)}{(3\rho^4 + 10\rho^3 - 4\rho^3\alpha + 8.0 + 14\rho^2 - 10\rho^2\alpha + 14\rho - 14\rho\alpha - 8\alpha)^2}\end{aligned}$$

Since $(3\rho^4 + 10\rho^3 - 4\rho^3\alpha + 8.0 + 14\rho^2 - 10\rho^2\alpha + 14\rho - 14\rho\alpha - 8\alpha)^2 > 0$ the roots of the numerator provides the extremum point of the function $l^L(\alpha,\rho)$. Rearranging the numerator, we have

$$\begin{aligned}G(\alpha,\rho) = &-33\alpha\rho^6 - \alpha(180 + 12\alpha)\rho^5 - \alpha(470 + 10\alpha + 8\alpha^2)\rho^4 - \alpha(604 + 20\alpha)\rho^3 \\ &- \alpha(300 + 104\alpha + 28\alpha^2)\rho^2 - 32\alpha(\alpha^2 + 4\alpha - 1)\rho + 64\alpha(1 - \alpha)\end{aligned}$$

First, note that $G(\alpha,\rho)$ is a continuous and differentiable polynomial in ρ. Inspecting the coefficients, we note that all coefficients are less than zero for $\rho^i \ \ \forall\, i = 2,\ldots,6$ and for $\alpha > 0$. The coefficient for ρ^0 is non negative for $\alpha \in [0,1]$. The coefficient for ρ is $-32\alpha(\alpha^2 + 4\alpha - 1)$ which is zero when $\alpha = (\sqrt{5} - 2)$ (and negative for $\alpha > (\sqrt{5} - 2)$, positive for $\alpha < (\sqrt{5} - 2)$). Therefore, inspecting the sequence of coefficients of the polynomial beginning from the highest degree, there is only one sign change for any $0 < \alpha < 1$. The sign changes from positive to negative when progressing from the coefficient of ρ to the coefficient of the constant term (ρ^0), when $(\sqrt{5} - 2) < \alpha < 1$ and changes sign from positive to negative ρ^2 to ρ when $\alpha < (\sqrt{5} - 2)$ and from ρ^2 to constant when $\alpha = (\sqrt{5} - 2)$. Applying Descartes' rule of signs, the polynomial $G(\alpha,\rho)$ has at most one root in the region $\rho \in (0,\infty)$. Therefore $l^L\left(\mathbf{1};\sigma^{\mathbf{L}}\right)$ has at most one extremum in $\rho \in (0,\infty)$. We have to show that the extremum always exists and that it is a maximum.

First, we note that $l^L(\alpha,\rho)$ is continuous and differentiable everywhere in the interval $[0,\infty)$. Consider the derivatives of the likelihood function at the extreme ends, $\lim_{\rho\to 0} \frac{dl^S(\mathbf{1},\sigma^{\mathbf{S}})}{d\rho}$ and $\lim_{\rho\to\infty} \frac{dl^S(\mathbf{1},\sigma^{\mathbf{S}})}{d\rho}$.

$$\begin{aligned}\lim_{\rho\to 0} \frac{dl^L(\alpha,\rho)}{d\rho} &= \lim_{\rho\to 0} \frac{\left[\begin{array}{c} -33\alpha\rho^6 - \alpha(180 + 12\alpha)\rho^5 - \alpha(470 + 10\alpha + 8\alpha^2)\rho^4 - \alpha(604 + 20\alpha)\rho^3 \\ -\alpha(300 + 104\alpha + 28\alpha^2)\rho^2 - 32\alpha(\alpha^2 + 4\alpha - 1)\rho + 64\alpha(1 - \alpha) \end{array}\right]}{(3\rho^4 + 10\rho^3 - 4\rho^3\alpha + 8 + 14\rho^2 - 10\rho^2\alpha + 14\rho - 14\rho\alpha - 8\alpha)^2} \\ &= \frac{64\alpha(1-\alpha)}{64(1-\alpha)^2} = \frac{\alpha}{1-\alpha} > 0 \ \ \forall \ \ \alpha \in (0,1]\end{aligned}$$

$$\lim_{\rho\to\infty}\frac{dl^L(\alpha,\rho)}{d\rho}=\lim_{\rho\to\infty}\frac{\left[\begin{array}{c}-33\alpha\rho^6-\alpha(180+12\alpha)\rho^5-\alpha(470+10\alpha+8\alpha^2)\rho^4-\alpha(604+20\alpha)\rho^3\\ -\alpha(300+104\alpha+28\alpha^2)\rho^2-32\alpha(\alpha^2+4\alpha-1)\rho+64\alpha(1-\alpha)\end{array}\right]}{(3\rho^4+10\rho^3-4\rho^3\alpha+8+14\rho^2-10\rho^2\alpha+14\rho-14\rho\alpha-8\alpha)^2}$$
$$<0\quad\forall\quad\alpha\in(0,1].$$

Since $\frac{dl^L(\alpha,\rho)}{d\rho}$ is continuous and differentiable in $\rho\in[0,\infty)$ we have $\frac{dl^L(\alpha,\rho)}{d\rho}=0$ at some $c\in(0,\infty)$. Therefore applying Rolle's Theorem, we show that there is at least one point where $\frac{dl^L(\alpha,\rho)}{d\rho}=0$. Hence, $l^L(\alpha,\rho)$ is unimodal in ρ over $[0,\infty)$ and reaches a maximum at some $\rho=\rho_{\max}(\alpha)\in(0,\infty)$.

4. First, let $D_l(\alpha,\rho)$ and $D_S(\alpha,\rho)$ be the denominators in the expressions of $l^L(\alpha,\rho), l^S(\alpha,\rho)$. We will consider the difference between the likelihood expressions.

$$\begin{aligned}D_S(\alpha,\rho)&=3\rho^4+13\rho^3-7\rho^3\alpha+24\rho^2-22\rho^2\alpha+2\alpha^2\rho^2+22\rho-22\alpha\rho+8-8\alpha\\&=(3\rho^4+10\rho^3-4\rho^3\alpha+14\rho^2-10\rho^2\alpha+14\rho(1-\alpha)+8(1-\alpha)+3\rho^3(1-\alpha)+2\rho^2(5-6\alpha+\alpha^2)\\&=D_L(\alpha,\rho)+3\rho^3(1-\alpha)+2\rho^2(5-6\alpha+\alpha^2)\end{aligned}$$

Now consider the likelihood ratios.

$$l^L(\alpha,\rho)=\frac{(3\rho^4+10\rho^3+7\rho^3\alpha+14\rho^2+20\rho^2\alpha+2\alpha^2\rho^2+14\rho+22\alpha\rho+8+8\alpha)}{(3\rho^4+10\rho^3-4\rho^3\alpha+8+14\rho^2-10\rho^2\alpha+14\rho-14\alpha\rho-8\alpha)}=\frac{N_L(\alpha,\rho)}{D_L(\alpha,\rho)}.$$
$$l^S(\alpha,\rho)=\frac{3\rho^4+13\rho^3+4\rho^3\alpha+8+24\rho^2+12\rho^2\alpha+22\rho+14\alpha\rho+8\alpha}{(3\rho^4+13\rho^3-7\rho^3\alpha+8+24\rho^2+2\alpha^2\rho^2-22\rho^2\alpha-22\alpha\rho+22\rho-8\alpha)}=\frac{N_S(\alpha,\rho)}{D_S(\alpha,\rho)}.$$
$$\begin{aligned}l^L(\alpha,\rho)-l^S(\alpha,\rho)&=\frac{\left[\begin{array}{c}-1\rho\alpha(21\rho^5\alpha-21\rho^5-160\rho^4-440\rho^3+160\rho^4\alpha+440\rho^3\alpha\\ +592\rho^2\alpha-592\rho^2-416\rho+416\rho\alpha-128+4\rho^3\alpha^2-4\rho^3\alpha^3+128\alpha)\end{array}\right]}{D_L(\alpha,\rho)D_S(\alpha,\rho)}\\&=\frac{\rho\alpha\left[\begin{array}{c}21\rho^5(1-\alpha)+160\rho^4(1-\alpha)+440\rho^3(1-\alpha)+592\rho^2(1-\alpha)\\ +416\rho(1-\alpha)+128(1-\alpha)+4\rho^3\alpha^2(1-\alpha)\end{array}\right]}{D_S(\alpha,\rho)D_L(\alpha,\rho)}\\&\geq 0\end{aligned}$$

For $l^L(\alpha,\rho)>1$, we need

$$\frac{(3\rho^4+10\rho^3+7\rho^3\alpha+8+20\rho^2\alpha+2\alpha^2\rho^2+14\rho^2+22\alpha\rho+14\rho+8\alpha)}{(3\rho^4+10\rho^3-4\rho^3\alpha+8+14\rho^2-10\rho^2\alpha+14\rho-14\alpha\rho-8\alpha)}>1$$
$$7\rho^3\alpha+20\rho^2\alpha+2\alpha^2\rho^2+22\alpha\rho>-4\rho^3\alpha-10\rho^2\alpha-14\alpha\rho \text{ which is true.}$$

Similarly we have $l^S(\alpha,\rho)>1\forall\alpha,\rho$.

Therefore $l^L(\alpha,\rho)\geq l^S(\alpha,\rho)\geq 1$ for all $\alpha\in[0,1]$. □

Appendix D: Strategies with Asymmetric buffers

In this section, we analyze herd behavior in queues with asymmetric buffers. There are α informed consumers in the market and the fraction $(1-\alpha)$ are consumers with signal strength $g \in [1/2, 1)$. Specifically, we examine small buffers with $N_1 = 2$ and $N_2 = 1$. Following the model specifications, the strategies at all states except $(0,0)$ and $(1,0)$ are immediately specified. We provide equilibrium results when consumers follow their signals at $(0,0)$. Note that Corollary 2 may not hold under asymmetric demand buffers. Nevertheless we noted, for our analysis, the less-informed consumers never joined any server with probability 1 (i.e. always join server 1). Due to brevity, we only present the results for consumers following their private signals at $(0,0)$.

Using the steps similar to Lemma C2, we write the steady state balance equations, and derive the likelihood ratios for different strategies at the state $(1,0)$. Using notations as before, let $l^L(\alpha,\rho,g), l^F(\alpha,\rho,g), l^S(\alpha,\rho,g)$ denote the likelihood ratios under the strategies $\sigma^L, \sigma^F, \sigma_S$ at the state.

We find $l^L(\alpha,\rho,g) = \frac{N_L(\alpha,\rho,g)}{D_L(\alpha,\rho,g)}$, $l^F(\alpha,\rho,g) = \frac{N_F(\alpha,\rho,g)}{D_F(\alpha,\rho,g)}$, and $l^S(\alpha,\rho,g) = \frac{N_S(\alpha,\rho,g)}{D_S(\alpha,\rho,g)}$ such that,

$$
\begin{aligned}
N_L(\alpha,\rho,g) &= (2\rho^2+2\rho+3\alpha\rho-3\rho g\alpha+3\rho g-4\alpha g+4\alpha+4g)(4\rho^4+9\rho^3+9\rho-2\rho^3\alpha+12\rho^2-2\rho^3 g \\
&\quad +2g\rho+\rho^5+2\rho^2g^2+\rho^3g^2-4\rho^2g^2\alpha-2\rho^3g^2\alpha+2\rho^2g^2\alpha^2+\rho^3g^2\alpha^2+\rho^3\alpha^2-2\rho^2\alpha \\
&\quad +2\rho^2\alpha^2+2\rho\alpha+4-4\rho^2g\alpha^2+6\rho^2g\alpha-2\rho^3\alpha^2g+4\rho^3\alpha g-2\rho^2g-2g\rho\alpha) \\
D_L(\alpha,\rho,g) &= (4\rho^4+8\rho^3+11\rho+12\rho^2-2g\rho+\rho^5+2\rho^2g^2+\rho^3g^2-4\rho^2g^2\alpha-2\rho^3g^2\alpha+2\rho^2g^2\alpha^2 \\
&\quad +\rho^3g^2\alpha^2+\rho^3\alpha^2-2\rho^2\alpha+2\rho^2\alpha^2-2\rho\alpha+4-4\rho^2g\alpha^2+6\rho^2g\alpha-2\rho^3\alpha^2g+2\rho^3\alpha g-2\rho^2g+2g\rho\alpha) \\
&\quad (2\rho^2+5\rho-3\alpha\rho+3\rho g\alpha-3\rho g-4\alpha+4\alpha g+4-4g) \\
N_F(\alpha,\rho,g) &= (-4g\alpha+4g+4\alpha+2\rho+3\alpha\rho-3\rho g\alpha+3\rho g+2\rho^2)(-2\alpha\rho^2+2\alpha\rho+4+9\rho+12\rho^2+9\rho^3+4\rho^4 \\
&\quad +\rho^5+4\alpha\rho^2g-2g\rho^2-1\rho^3g-2\rho^3\alpha+2\rho^3g\alpha+\alpha^2\rho^3-1\alpha^2\rho^3g+2\alpha^2\rho^2-2\alpha^2\rho^2g) \\
D_F(\alpha,\rho,g) &= (2\alpha\rho^2+\rho^3g+4+9\rho+10\rho^2+8\rho^3+4\rho^4+\rho^5-2\alpha\rho^2g-1\rho^3g\alpha+2g\rho^2+\rho^3\alpha) \\
&\quad (-4\alpha+4g\alpha+4-4g+5\rho-3\alpha\rho+3\rho g\alpha-3\rho g+2\rho^2) \\
N_S(\alpha,\rho,g) &= (-4g\alpha+4g+4\alpha-2\rho-3\alpha\rho-3\rho g\alpha+3\rho g+2\rho^2)(4+11\rho+12\rho^2+8\rho^3+4\rho^4+\rho^5) \\
D_S(\alpha,\rho,g) &= (\alpha^2\rho^3-2\alpha\rho+2\alpha^2\rho^2-2\alpha\rho^2+4+11\rho+12\rho^2+8\rho^3+4\rho^4 \\
&\quad +\rho^5-\alpha^2\rho^3g+\rho^3g\alpha-2\alpha^2\rho^2g+2\alpha\rho^2g)(-4\alpha+4g\alpha+4-4g+5\rho-3\alpha\rho+3\rho g\alpha-3\rho g+2\rho^2)
\end{aligned}
$$

Following the same process as in Lemma C2, we can show that $l^L(\alpha,\rho,g)$ is decreasing in ρ and as $\lim_{\rho\to 0} l^L(\alpha,\rho,g) = \frac{1+\alpha}{1-\alpha}$. So, many of the properties of the likelihood ratio curves derived for the symmetric case continue to hold. In particular, we can show that for any given g, the likelihood ratio l^L is decreasing in ρ. This indicates for $c\tau \le \alpha\Delta$, there exist $\overline{\rho}$, so that joining the longer queue always is an equilibrium strategy when $\rho \in (0, \overline{\rho})$. This property is illustrated in Figure D1. Also, note from Figure D1, longer queue joining strategy is in equilibrium for low arrival rates.

Assuming Independent Queues: Asymmetric Case: Let N_1, N_2 be the buffer sizes, and let

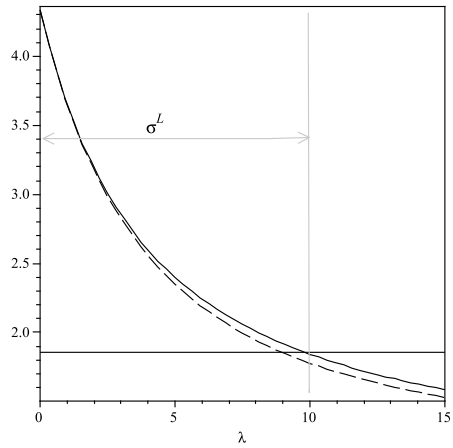

Figure D1: *Likelihood ratio curves l^L, l^S (under strategies $\sigma^{\mathbf{S}}$ and $\sigma^{\mathbf{L}}$ respectively) are represented by thick, dotted curves respectively. The threshold $\frac{\Delta+c\tau}{\Delta-c\tau}$ is indicated by the thick horizontal line. The longer-queue joining equilibrium strategy is shown by arrow marks. The heterogenous market with parameters $\Delta = 1, c = 0.3, \alpha = 0.25$ and $g = 0.75$ is shown. Notice that even for asymmetric buffers, joining the longer queue σ^L is in equilibrium for low arrival rates, and at higher arrival rates, consumers join the shorter queue.*

$N_1 > N_2$ Then we have,

$$\begin{aligned} q_{i,j}(n,\sigma') &= \frac{(g'\rho)^{n_1}(1-g'\rho)}{(1-(g'\rho)^{N_i+1})} \quad j=i \\ &= \frac{((1-g')\rho)^{n_2}(1-(1-g')\rho)}{(1-((1-g')\rho)^{N_i+1})} \quad j\neq i \end{aligned}$$

Without loss of generality let $N_1 \geq N_2$. Then for all $(n_1, n_2) \in \mathcal{N}$,

$$\begin{aligned} \pi_1(n_1,n_2) &= q_{1,1}(n_1,\sigma')\cdot q_{2,1}(n_2,\sigma') \\ &= \frac{(g'\rho)^{n_1}(1-g'\rho)}{(1-(g'\rho)^{N_1+1})}\frac{((1-g')\rho)^{n_2}(1-(1-g')\rho)}{(1-((1-g')\rho)^{N_2+1})} \\ \pi_2(n_1,n_2) &= q_{1,2}(n_1,\sigma')\cdot q_{2,2}(n_2,\sigma') \\ &= \frac{((1-g')\rho)^{n_1}(1-(1-g')\rho)}{(1-((1-g')\rho)^{N_1+1})}\frac{(g'\rho)^{n_2}(1-g'\rho)}{(1-(g'\rho)^{N_2+1})} \\ l^{asym}(n_1,n_2) &= \frac{\pi_1(n_1,n_2)}{\pi_2(n_1,n_2)} = (\frac{g'}{1-g'})^{n_1-n_2} \\ &= (\frac{\alpha+g(1-\alpha)}{(1-g)(1-\alpha)})^{n_1-n_2}\left[\frac{(1-((1-g')\rho)^{N_1+1})(1-(g'\rho)^{N_2+1})}{(1-(g'\rho)^{N_1+1})1-(1-((1-g')\rho)^{N_2+1})}\right] \\ &> (\frac{\alpha+g(1-\alpha)}{(1-g)(1-\alpha)})^{n_1-n_2}. \end{aligned}$$

Therefore, we can show that herding occurs at more states (conditional on $n_1 > n_2$) given $N_1 > N_2$. The threshold b_1^{asym} that satisfies $(l^{asym}(n_1,n_2))^{b_1^{asym}} = \frac{g(\Delta+b_1c\tau)}{(1-g)(\Delta-b_1c\tau)}$ is such that $b_1^{asym} < b_1$ when $n_1 > n_2$, and $b_1^{asym} < b_1$ when $n_2 > n_1$.

Appendix E: Markets with Partial Observability

We now analyze the herd behavior in partially observable markets. In our two-server market, one queue length is observed, and the other queue remains unobserved. This might occur in the case of consumers observing the congestion at the restaurant they pass first. They might use the congestion at the restaurant in deciding to dine there, or decide to go to another restaurant down the street. Suppose that all consumers arriving at the market observed the queue length at server 1, while the queue at server 2 remains unobserved. Some consumers decide to join queue 1 depending on the length of the queue, and the rest balk from queue 1 and join the other queue that is unobservable.

Proposition E5. *When consumers minimize the worst case regret, they join the observable queue when its length, n is less than a threshold $\hat{n}$ such that $\rho^{\hat{n}+2} = \frac{\hat{n}}{\hat{n}+1}$, and join the unobservable queue otherwise.*

Proof. of **Proposition E5:**

Suppose a consumer arrives at the market and observes the queue length to be n, and assumes (rationally) the equilibrium queue arrival rate at the other unobservable queue is λ_e. Let 1 be the observable queue. The consumer minimizes regret instead of maximizing expected utility. Let $R(i|n_1, s) \forall\ i, s \in \{1, 2\}, n_1 \in \mathbb{N}$ represent the worst case regret of the consumer with signal s who joins queue i, when she observes the queue length to be n_1. Let $\rho_e = \lambda_e \tau$.
Therefore, $\max R(1|n_1, s, V_1 > V_2) = \max\{0, (V_+ - c\tau/(1-\rho_e)) - (V_- - (n_1+1)c\tau)\}$.
Similarly, $\max R(2|n_1, s, V_1 > V_2) = \max\{0, (V_+ - (n_1+1)c\tau) - (V_- - c\tau/(1-\rho_e))\}$. Again, let $a^{mr}(s, n) \in \{1, 2\}$ be the action that minimizes maximum regret at a state n when seeing signal s.

$$
\begin{aligned}
a^{mr}(s, n_1) &= \arg\min_{\{1,2\}} \{\max R(1), \max R(2)\} \\
&= \arg\min_{\{1,2\}} \{\max\{0, (V_+ - c\tau/(1-\rho_e)) - (V_- - (n_1+1)c\tau)\}, \\
&\quad \max\{0, (V_+ - (n_1+1)c\tau) - (V_- - c\tau/(1-\rho_e))\}\} \\
&= \arg\min_{\{1,2\}} \{(V_+ - V_-) - (c\tau/(1-\rho_e) - (n_1+1)c\tau), (V_+ - V_-) - ((n_1+1)c\tau - (c\tau/(1-\rho_e)))\} \\
&= \arg\min_{\{1,2\}} \{-(c\tau/(1-\rho_e) - (n_1+1)c\tau), -((n_1+1)c\tau - c\tau/(1-\rho_e))\}. \\
&= \arg\max_{\{1,2\}} \{(c\tau/(1-\rho_e) - (n_1+1)c\tau), ((n_1+1)c\tau - c\tau/(1-\rho_e))\}.
\end{aligned}
$$

This implies that the consumer chooses to join queue 1, if $n_1 \leq \hat{n} = (1/(1-\rho_e)) - 1$, otherwise she joins queue 2, which has expected equilibrium arrival rate of λ_e (regardless of the signal). We assume that the indifferent consumer chooses the observable queue.
We now solve for the equilibrium arrival rate λ_e. Note that the volume of consumers arriving to queue 2 is identical to the rate of consumers who balked from queue 1. Therefore $\lambda_e = \lambda \Pr[n_1 =$

$\hat{n}+1]$. Plugging in the expression for $\hat{n}$ and using $M/M/1/N$ queue expressions, we have

$$\begin{aligned}
\lambda_e &= \lambda \frac{\rho^{\hat{n}+1}}{1-\rho^{\hat{n}+2}}. \\
\hat{n} &= \frac{1}{1-(\rho\frac{\rho^{\hat{n}+1}}{1-\rho^{\hat{n}+2}})} - 1 \\
&= \frac{1-\rho^{\hat{n}+2}}{1-2\rho^{\hat{n}+2}} - 1 \\
\Rightarrow \rho^{\hat{n}+2} &= \frac{\hat{n}}{\hat{n}+1}.
\end{aligned}$$

□

Proposition E5 indicates that consumers join the first server below a threshold queue length, based on the expected equilibrium queue length at the unobservable queue. The worst case regret occurs when consumers join a long queue at server 1 and then find out ex post that server 2 is the better server, and was also not as congested as the first server when they arrived. In general, it appears that when consumers minimize worst case regret, the herd behavior observed under rational consumer decision making persists, but may not be as pronounced.