The Value of Congestion

Laurens G. Debo †
Tepper School of Business
Carnegie Mellon University
Pittsburgh, PA 15213

Christine A. Parlour †
Tepper School of Business
Carnegie Mellon University
and Haas School of Business
UC Berkeley

Uday Rajan §
Ross School of Business
University of Michigan
Ann Arbor, MI 48109.

February 13, 2006

Abstract

We provide a model in which a queue for a good communicates the quality of the
good to consumers. Agents arrive randomly at a market, and observe the queue length
and a private signal (good or bad) about the good. Service departures from the queue
are also random. Agents decide whether to join the queue and obtain the good or to
balk. When waiting costs are zero, agents receiving a bad signal join the queue only if
it is long enough. When the waiting costs are non-zero, agents do not join the queue if
it is too long. Furthermore, agents with bad signals may play non-threshold strategies.
In equilibrium, an agent is more likely to enter a queue for a low quality good than balk
from a queue for a high quality good. Under specified conditions, both a high quality
firm and a social planner maximizing consumer surplus prefer a low service rate to a
high one, due to the information externality.

*We thank seminar participants at Columbia, CETC and the Carnegie Mellon/Technische Univer-
siteit Eindhoven Collaborative Workshop (2005). The current version of this paper is maintained at
†Tel: (412) 268-8403, E-mail: laurdebo@andrew.cmu.edu
‡Tel: (412) 268-5806, E-mail: parlourc@andrew.cmu.edu or parlour@haas.berkeley.edu
§Tel: (734) 764-2310, E-mail: urajan@umich.edu
1 Introduction

Customers frequently wait before they can consume a good. Lines outside nightclubs, rides at amusement parks and waiting lists for new products are part of everyone’s experience. Further, industrial firms frequently announce their backlog of orders: they make other customers aware of a queue for their product. Becker (1991) posits that goods such as sporting matches or restaurant meals are more valuable if many others are also consuming them. As people enjoying consuming together, congestion increases the value of these goods. His explanation clearly does not extend to less convivial goods for which waiting is merely inconvenient and therefore costly. For such goods, what benefit does a firm obtain from displaying the number of consumers who are waiting? If consumers are uncertain about the quality of a product, a queue provides credible information to later consumers about the beliefs of earlier arrivals. The firm faces a tradeoff between congestion or waiting costs incurred by consumers and the positive information externality that customers waiting generate for others when they choose to wait.

We model this tradeoff in a queueing model in which a firm can be of either high or low quality. Risk neutral consumers arrive sequentially at a market. Each consumer receives a binary signal of the quality of the firm and must decide whether to purchase the good. Purchasing entails joining a “first-come, first-served” queue and waiting. There may be a cost associated with waiting. In the absence of waiting costs, the queue exerts a pure informational externality: an arriving consumer updates her beliefs about the firm on seeing the length of the queue. When there is a cost to waiting, the queue also imposes a negative congestion externality on new arrivals.

Consumers who have joined the queue leave once they have been serviced. As is standard in queueing models, both arrival and service are random events modeled as Poisson processes. Stochastic arrival and departure imply that the length of the queue varies over time. We assume that an arriving consumer does not observe the entire history of the game and so does not know how many people arrived before her. Instead, she sees the number of people currently in the queue. Her decision to join the queue, of course, is endogenous.

The inference problem faced by an agent is potentially complicated. On arriving, she observes the queue but not the queue length faced by agents who arrived before her. Thus, in principle he forms beliefs about the queue length faced by every queue member and

\footnote{See, for example, PR Newswire, New York, Aug 19, 1999 “Gerber Scientific, Inc. Reports First Quarter EPS Up 17%; Orders and Backlog Rise on Strong New Product Introductions.”}
the signal that each agent received. To solve for equilibrium, we make use of a result from queueing theory, namely that a randomly arriving agent faces a state drawn from the steady state distribution. Thus, as firms of different qualities give rise to different steady state distributions, the queue length subsumes all beliefs.

In equilibrium, agents with good signals join the queue as long as it is not too long. Beyond some threshold, the waiting costs imply that joining the queue is not worthwhile. If waiting costs are zero, this threshold is infinite. Further, in this case, agents with bad signals join the queue if it is long enough. This contrasts with standard queuing models, in which long queues deter entry. If there are positive waiting costs, we demonstrate the existence of action “holes”: queue lengths for which agents with bad signals will not join the queue. If the queue crosses such a threshold, it must be due to the arrival of some agent with a good signal. Thus, the existence of waiting costs sometimes makes it easier for the public to infer an agent’s private signal if she joins the queue.

An agent who observes a longer queue believes that the firm is more likely be of high quality. Further, the steady-state distribution of queue lengths for high quality firms first order stochastically dominates that of low quality firms. Thus, waiting times are, on average, longer for firms with high quality.

To capitalize on the information externality, firms have an incentive to maintain long queues by reducing the service rate. Similarly, a social planner who seeks to maximize consumer surplus also has an incentive to keep the service rate low. Since the firm and planner have different goals, they will choose different service rates. Nevertheless, we demonstrate that, for some parameter values, both a firm and the planner would prefer not to increase the service rate, even if doing so is costless. That is, the benefits of congestion, in terms of the additional information about quality provided to consumers, outweigh the effects of the extra waiting costs borne by consumers.

When agents with good and bad signals play the same strategies at a given queue length, we say they display herd behavior. The canonical herding models are due to Bikhchandani, Hirshleifer and Welch (1992) and Banerjee (1992). In these models, agents observe the entire history of the game, in terms of actions taken by previously arriving agents. Each agent has a private signal, and infers the quality of a good after viewing the actions of prior agents. In our model, agents observe the queue when they arrive, but have no other public information, not even how many agents have arrived previously. It is natural to consider that an agent arriving at the market has no information about how many agents arrived previously and decided not to consume the good, or indeed consumed the good and left the

\footnote{In particular, queueing, observing only the length of the queue, and waiting costs appear to resonate well with the restaurant example of Banerjee (1992).}
market. Instead, an agent observes only those agents currently waiting for the good.

At any point of time, the random arrival and departure rates ensure that queue lengths change for newly arriving agents. Hence, there cannot be a cascade—a point in time beyond which all agents take the same action. Starting with a long queue, if a lot of agents leave the queue because they were serviced, the queue may dwindle before the next agent comes along. In other words, empty queues appear with positive probability, and would do so even if all agents entered. But we show that there is no herding at the empty queue—an agent with a bad signal takes a different action than an agent with a good signal. Therefore, we have little to say on informational cascades. Nevertheless, since the length of the queue communicates information about quality, our model can be interpreted in terms of social learning. A comprehensive summary of herding and social learning appears in Chamley (2004).

Our model is also related to the literature on “word of mouth communication” and social learning, in which agents receive samples of the history. Banerjee (1993) presents a model in which agents optimally choose whether to pass on rumors. He finds that not all investors with good signals will invest. Ellison and Fudenberg (1993, 1995) find that with boundedly rational agents, word of mouth communication can lead all agents to a superior outcome, even if each individual receives little social information. If each agent observes only her predecessor’s action, beliefs and actions can cycle forever, with longer periods of uniform behavior and rare switches (Celen and Kariv, 2004). Smith and Sorensen (1997) consider a sequential action model in which agents receive a random sample of the history. They find that social learning does not converge to the truth if beliefs are bounded (as they are in our case).

The queueing model provides a natural way to restrict the histories that an agent sees. In particular, the actions of other agents are observed only if they chose to purchase the product; the decisions of agents who did not join the queue are not observed. In addition, as the service rate is random, an arriving agent does not know how long the queue was when any agent currently in the queue chose to join. Further, as a result of the stochastic service process, the queue length changes over time even when no one else joins. Therefore, agents arriving at different times observe different histories, and may come to different conclusions about the quality of the product, even if they had similar private information about quality.

Finally, our model builds upon the extensive literature on queueing and games. Optimal queue lengths are examined by Naor (1969) and Hassin (1985). Hassin (1986) considers the effect of revealing information about queue length, and therefore waiting times when there are congestion externalities. A comprehensive review of the economic aspects of queueing appears in Hassin and Haviv (2003).
2 Model

A firm sells an experience good which can be of high \((h)\) or low quality \((\ell)\). The utility an agent gets from the good is \(v \in \{v_\ell, v_h\}\), where \(v_\ell < 0 < v_h\). This utility is assumed to be net of the price charged by the firm, which is not explicitly modeled. While the firm knows the quality of its own good, it cannot credibly communicate it to consumers.

It takes time for the firm to service (i.e., provide the good to) each consumer. The firm’s service time for each consumer is exponentially distributed with parameter \(\mu\), and is independent across consumers. Consumers suffer disutility from waiting, with a waiting cost of \(c \geq 0\) per unit time.

Consumers are risk neutral, and arrive sequentially at the market according to a Poisson process with parameter \(\lambda\). If agents arrive faster than they are serviced, they queue. The queue is served on a first-come first-served basis. Agents’ prior belief that the product is high quality is \(p_0\). In addition, each agent receives a private signal \(s \in S = \{g, b\}\) about the quality of the good, where \(\Pr(s = g \mid v = v_h) = \Pr(s = b \mid v = v_\ell) = q \in \left(\frac{1}{2}, 1\right)\). Agents know neither the order in which they arrive at the market, nor the history of actions taken by previously arriving agents. However, each agent does observe the length of the queue when he arrives at the market.

The agent takes an action \(a \in \{\text{join, balk}\}\), where \(a = \text{join}\) indicates a decision to acquire the product, or join the queue. Once he joins the queue, he may not renege; i.e., he cannot leave until he has been served. Joining the queue is therefore synonymous with consuming the product. If he chooses to balk (i.e., not acquire the good), he obtains a reservation utility of zero.

A mixed strategy for an agent is a mapping \(\alpha : S \times N \to [0, 1]^{S \times N}\), where \(\alpha(s, n)\) denotes the probability that an agent with signal \(s\) who sees queue length \(n\) joins the queue and \(N\) is the set of non-negative integers. We consider stationary Markov perfect-Bayesian equilibria of this game. The state space is defined by \(S \times N\). The type of an agent is completely defined by \((s, n)\). Since an agent has no other distinguishing feature, the equilibrium we consider is perforce symmetric.

Consider the problem faced by an agent who observes state \((s, n)\). Though she knows the number of people currently in the queue, she does not know the state of the system these faced when they chose to join the queue. Her inference problem is potentially complicated, since she may have beliefs over the state of the system when each of these agents joined, as well as beliefs over the number of agents who chose not to join the queue.

Consider a randomly arriving agent. Suppose all other agents are playing the strategy \(\alpha\). Then, given signal, \(s\), and queue length, \(n\), the randomly arriving agent’s posterior belief
that the product is of high quality is denoted as $\theta_s(n; \alpha)$. Due to the memoryless property of the exponential distribution, the expected time to serve each agent in the queue (including the one currently being served) is $\frac{1}{\mu}$. Thus, for the newly arrived agent, the expected total waiting time before he leaves the system is $(n + 1)\frac{1}{\mu}$. Given his signal and observed queue length $(s, n)$, the agent’s expected net utility from joining the queue is

$$u(s, n, \alpha) = v_\ell + \theta_s(n; \alpha)(v_h - v_\ell) - \left(\frac{n + 1}{\mu}\right)c.$$

(1)

The agent will join the queue with positive probability only if $u(s, n, \alpha) \geq 0$.

We make the following assumptions about the parameters. Let $p_s = \text{Prob}(h \mid s)$. Then, from Bayes’ rule, $p_g = \frac{p_0 q}{p_0 q + (1 - p_0)(1 - q)}$ and $p_b = \frac{p_0 (1 - q)}{p_0 (1 - q) + (1 - p_0) q}$.

**Assumption 1**

(i) Either $c > 0$ or $\frac{\lambda}{\mu} < 1$.

(ii) $p_g v_h + (1 - p_g) v_\ell > \frac{c}{\mu} > p_b v_h + (1 - p_b) v_\ell$

Part (i) ensures that the system is stationary, i.e. the expected queue length remains finite. If there are congestion costs (so that $c > 0$), even if all agents believe the good to be of high quality there is a maximum queue length. If there are no congestion costs, then the arrival rate of agents to the market is less than the service rate $\mu$. Part (ii) states that an agent who acts only on the basis of her own signal, and ignores any information in the observed queue length, joins an empty queue if and only if her signal is good. This is similar to the usual assumption in the cascades literature that an agent who ignores the information provided in other agents’ actions will acquire the product if she has a good signal, but not if she has a bad one.

In a symmetric equilibrium, the agent’s expected payoff in state $(s, n)$ from playing the strategy $\alpha$ is $\alpha(s, n)u(s, n, \alpha)$. Hence, $\alpha$ defines a perfect Bayesian equilibrium if it maximizes this expected payoff in each state $(s, n)$.

**Definition 1** A strategy $\alpha$ is a stationary Markov perfect Bayesian equilibrium if, for each $s \in \{g, b\}$ and each $n \in N$,

$$\alpha(s, n) \in \arg \max_{x \in [0, 1]} xu(s, n, \alpha),$$

where $u(s, n, \alpha)$ is defined by (1), and $\theta_s(n; \alpha)$ is defined by Bayes’ rule whenever feasible.

2.1 Posterior beliefs

Recall that $p_h$ is the posterior probability that the firm has high quality, if an agent observes signal $s$ but not the queue length. Let $\pi_j(n; \alpha)$, for $j = h, \ell$, be the probability that a
randomly arriving agent observes \( n \) people in the queue, given that the true quality of the firm is \( j \) and all agents play the strategy \( \alpha \). Then, the agent’s posterior belief that the firm has high quality, after having observed signal \( s \) and queue length \( n \) is

\[
\theta_s(n; \alpha) = \frac{p_s \pi_h(n; \alpha)}{p_s \pi_h(n; \alpha) + (1 - p_s) \pi_\ell(n; \alpha)} \quad \text{if } \pi_j(n; \alpha) > 0 \text{ for some } j \in \{h, \ell\}. \tag{2}
\]

As usual, perfect Bayesian equilibrium places no restrictions on belief \( \theta_s(n, \alpha) \) if \( \pi_j(n; \alpha) = 0 \) for both \( j = h \) and \( j = \ell \).

To characterize equilibrium, we exploit a result from the queuing literature: the PASTA (Poisson Arrivals See Time Averages) property. A general formalization by Wolff (1982) demonstrates that a randomly arriving consumer sees a queue length sampled from the stationary distribution. Let \( \pi_j(n; \alpha) \) be the stationary distribution when the quality of the firm is \( j \in \{\ell, h\} \) and all agents play strategy \( \alpha \). We derive the stationary distribution as follows. Let \( r_j(n; \alpha) \) be the ex ante probability (i.e., given a queue length \( n \), but before the signal \( s \) is observed) that an agent who plays strategy \( \alpha \) joins a queue that already has \( n \) agents, given that the firm’s quality is \( j \). Then,

\[
r_j(n; \alpha) = \begin{cases} q\alpha(g, n) + (1 - q)\alpha(b, n) & \text{if } j = h \\ (1 - q)\alpha(g, n) + q\alpha(b, n) & \text{if } j = \ell \end{cases} \tag{3}
\]

Now, if there are \( n \) agents in the queue, the rate at which a new agent joins the queue is \( \lambda r_j(n; \alpha) \). The “joining rate” depends both on the true quality of the firm, \( j \in \{\ell, h\} \), and on the strategy being played by the agents, \( \alpha \). As the service rate is constant and equal to \( \mu \), for any \((j, \alpha)\) pair, the induced queuing system is a birth and death process.

Consider a queue length of \( n \). There are only two ways in which this could have occurred. First, there could have been a queue of \( n + 1 \) people and with rate \( \mu \) the agent at the front of the queue was served, or the queue could have been of length \( n - 1 \) and with rate \( \lambda r_j(n; \alpha) \) a new agent could have joined. Thus, given quality \( j \) and strategy \( \alpha \), the expected number of entrances into state \( n > 0 \) per unit of time is \( \pi_j(n + 1; \alpha)\mu + \pi_j(n - 1; \alpha) \lambda r_j(n; \alpha) \). Conversely, if the queue is at \( n \), it can either increase due to an arriving agent or decrease due to a service completion. The expected number of departures from state \( n \) per unit of time is \( \pi_j(n; \alpha)(\lambda r_j(n; \alpha) + \mu) \). In steady state, the expected number of entrances into any state \( n \) is equal to the expected number of departures from state \( n \), generating a flow balance equation for each \( n \). The stationary probabilities are the solution to all such flow balance equations. With \((i)\) of Assumption 1, the solution is unique. The following Lemma characterizes two properties of the solution.

**Lemma 1** Suppose all agents follow the strategy profile \( \alpha \). Then, for \( j = h, \ell \), the stationary
probability of observing a queue of length \( n \) satisfies

\[
\pi_j(n; \alpha) = \pi_j(0; \alpha) \left( \frac{\lambda}{\mu} \right)^n \prod_{k=0}^{n-1} r_j(k; \alpha),
\]
with

\[
\pi_j(0; \alpha) = \frac{1}{1 + \sum_{n=1}^{\infty} \left( \frac{\lambda}{\mu} \right)^n \prod_{k=0}^{n-1} r_j(k; \alpha)}.
\]

Notice that \( \pi_j(n+1; \alpha) = \pi_j(n; \alpha) \frac{\lambda}{\mu} r_j(n; \alpha) \). Therefore, the stationary probability of observing \( (n+1) \) agents in the queue is zero if and only if \( r_j(n; \alpha) = 0 \); i.e., if no signal will induce an agent to join a queue that already has \( n \) agents. Further, if \( \pi_j(n; \alpha) = 0 \) for any \( n \), then \( \pi_j(k; \alpha) = 0 \) for all \( k \geq n \), and the queue cannot grow beyond length \( n \).

### 2.2 Uniqueness

Existence of a perfect Bayesian equilibrium follows from standard results.\(^3\) However, while the equilibrium exists, it is not necessarily unique. The information content in some queues may induce more agents to join the queue, while congestion costs encourage agents to balk. As the information content depends on the equilibrium actions of other agents, multiple equilibria can arise.

**Example 1** Let \( p_0 = 0.5 \) and \( q = 0.575 \). Suppose \( \lambda = 1 \), \( \mu = 1.3 \), \( v_h = 1.1 \), \( v_l = -0.9 \), and \( c = 0 \). Then, the following equilibria are obtained, where \( \alpha^* \) denotes the equilibrium strategy:

1. \( \alpha^*(b, 0) = 0 \), and \( \alpha^*(b, n) = 1 \) for all \( n \geq 1 \). So, an agent with a bad signal joins if and only if there is at least one person already in the queue. For each \( s, n \), the equilibrium action \( \alpha^*(s, n) \) is a strict best response.

2. \( \alpha^*(b, 0) = \alpha^*(b, 1) = 0 \), and \( \alpha^*(b, n) = 1 \) for all \( n \geq 2 \). So, the agent with a bad signal joins if and only if there are at least two people already in the queue. For each \( s, n \), the equilibrium action \( \alpha^*(s, n) \) is a strict best response.

3. \( \alpha^*(b, 0) = 0 \), \( \alpha^*(b, 1) = 0.735 \), and \( \alpha^*(b, n) = 1 \) for all \( n \geq 2 \). An agent with a bad signal joins if there are at least two queue people already in the queue, and mixes between joining and not joining when there is only one person in the queue.

\(^3\)For example, Theorem 5.4 of Rieder (1979).
2.3 Best responses

Suppose all other agents are playing the strategy $\alpha$, and consider the optimal action of an agent who, upon arrival, observes signal $s$ and is faced with a queue length $n$. The agent is willing to join the queue only if his expected payoff $u(s, n, \alpha)$ is non-negative, or, in other words, if the posterior belief $\theta_s(n; \alpha)$ that the firm is of high quality is sufficiently high. It is useful to write this condition in terms of the stationary probabilities over queue lengths.

Let $\psi_s(n; \alpha) = \left(\frac{1-p_s}{p_s}\right) \left(\frac{\pi_\ell(n; \alpha)}{\pi_h(n; \alpha)}\right)$. This is the overall likelihood that the good is of low quality, given the prior, the signal $s$, and the queue length $n$. Further, let $\gamma(n) = -\frac{v_h-(n+1)c}{v_r-(n+1)c}$. Thus, $\gamma(n)$ is the ratio of the utility from consuming a high quality good to the utility of consuming a low quality good, when the queue already has $n$ consumers. Notice that $\gamma(0) > 0$, since $v_\ell < 0$. For $c > 0$, $\gamma$ is a strictly decreasing function of $n$, with $\gamma(n) \rightarrow -1$ as $n \rightarrow \infty$.

**Lemma 2** Suppose all other agents are playing the strategy $\alpha$, and queue length $n$ has positive probability. Then, the best responses of a randomly arriving agent who faces state $(s, n)$ are given by: $\alpha^*(s, n) = 0$ if $\psi_s(n; \alpha) > \gamma(n)$, $\alpha^*(s, n) \in [0, 1]$ if $\psi_s(n; \alpha) = \gamma(n)$, and $\alpha^*(s, n) = 1$ if $\psi_s(n; \alpha) < \gamma(n)$.

That is, for an agent to join a queue, the stationary probability induced by the high quality firm (i.e., the firm that produces a high quality good) at that queue length must be sufficiently high, compared to the corresponding probability induced by the low quality firm (i.e., a firm that produces a low quality good). The condition in Lemma 2 can be interpreted as follows. An agent is willing to enter the queue if $\psi_s(n; \alpha) \leq \gamma(n)$. The left-hand side is the overall likelihood the firm has low quality. The right-hand side, $\gamma(n)$, is inversely related to the congestion cost imposed by a queue of length $n$. The condition in Lemma 2, therefore, says that the likelihood the firm is of low quality must be sufficiently small for an agent to join the queue. Any stationary Markov perfect Bayesian equilibrium is a fixed point of the best response correspondence exhibited in Lemma 2.

Going forward, for notational brevity, we often suppress the dependence of $\pi, r$, and $\psi$ on $\alpha$.

3 Properties of Equilibrium

We now characterize equilibrium properties. As Example 1 demonstrates, for some parameter values, there are multiple equilibria. Nevertheless, the properties in this section are common to all equilibria. Given the structure of the posterior belief in equation (2), it is immediate that, in any state in which an agent with a bad signal joins the queue, so will
an agent with a good signal. Further, high quality firms are more likely to generate good signals. Finally, observe that if an agent with a good signal does not join the queue at some length \( n \), the queue can never grow beyond \( n \), since an agent with a bad signal will not join either.

**Lemma 3** Let \( \alpha^* \) denote an equilibrium strategy. Then,

(i) the probability an agent joines the queue is higher when the agent has a good signal and when the firm has high quality; i.e., for all \( n \in \mathbb{N} \), \( \alpha^*(g,n) \geq \alpha^*(b,n) \) and \( r_h(n;\alpha^*) \geq r_\ell(n;\alpha^*) \).

(ii) for any queue length \( n \), if \( \alpha^*(g,n) = 0 \), then \( \pi_\ell(\tilde{n};\alpha^*) = \pi_h(\tilde{n};\alpha^*) = 0 \) for all \( \tilde{n} \geq n+1 \).

Given equation (4), the stationary probabilities of a high and low quality firm at a queue length of zero, and thus the equilibrium strategy at zero \( (\alpha^*(s,0)) \) are an important characteristic of the equilibrium. The first part of Proposition 1 establishes that a low quality firm is more likely to have an empty queue than a high quality firm. Further, an agent with a bad signal does not join an empty queue. By contrast, an agent with a good signal joins an empty queue with positive probability.

**Proposition 1** In equilibrium,

(i) a queue length of zero is more likely if the firm is of low quality, or \( \pi_\ell(0;\alpha^*) > \pi_h(0;\alpha^*) \).

(ii) an agent with a bad signal never joins the empty queue, whereas an agent with a good signal joins the empty queue with positive probability; i.e., \( \alpha^*(b,0) = 0 \) and \( \alpha^*(g,0) > 0 \).

It follows immediately from Proposition 1 that from a queue length of zero, a length of one can only be reached if an agent with a good signal joins. In general, as costs are independent of the signals, observing a longer queue length is “good news.” By contrast, a queue length of zero is “bad news” for an agent: the likelihood that the firm has low quality is higher. With no information about queue length, an agent believes with probability \( p_s \) that the firm has high quality. Once she sees a queue length of zero, she revises this belief downward.

Further, Proposition 1 and Lemma 3 imply that agents with good signals play “threshold” strategies: if the waiting cost is strictly positive, there is a threshold \( \bar{n}(c) \) such that an agent with signal \( g \) joins the queue with positive probability until the queue reaches \( \bar{n}(c) \). At this point, the agent does not join the queue. From Proposition 1, neither does an agent
with signal $b$. Hence, $\bar{n}(c)$ is the maximal queue length observed in equilibrium. If the queue length is $\frac{v_h}{c} \mu$, an agent who knows the good has high quality finds the waiting costs too high to join the queue. Thus, when $c > 0$, it is immediate that $\bar{n}(c) \leq \frac{v_h}{c} \mu$. When $c = 0$, we define $\bar{n}(0) = \infty$.

Let $W_j(\alpha^*; c)$, for $j = h, \ell$, be the expected waiting time for a randomly arriving customer at a firm of quality $v_j$. From the PASTA property, randomly sampled customers face the stationary distribution. Thus,

$$W_j(\alpha^*; c) = \sum_{n=0}^{\bar{n}(c)} \pi_j(n; \alpha^*) r_j(n; \alpha^*) \frac{n + 1}{\mu}, \text{ for } j = h, \ell.$$ 

The queue length distribution induced by a high quality firm first order stochastically dominates the distribution induced by a low quality firm, so that on average a new consumer observes a longer queue at the high quality firm. Given that the average queue length is longer for the high quality firm, it follows that the equilibrium expected waiting time is higher also for the high quality firm.

**Proposition 2** In equilibrium,

(i) For any $n, \tilde{n}$ such that $n \leq \tilde{n} \leq \bar{n}(c)$, $\frac{\pi_h(\tilde{n}; \alpha^*)}{\pi_h(n; \alpha^*)} \leq \frac{\pi_h(n; \alpha^*)}{\pi_h(n; \alpha^*)}$, or the likelihood ratio is weakly decreasing in queue length. (ii) the expected queue length is higher for a high quality firm, and (iii) the expected waiting time for a high quality product is greater than the expected waiting time for a low quality product, or $W_h(\alpha^*; c) > W_\ell(\alpha^*; c)$.

Queues, therefore, are physical manifestations of a firm’s quality. While high quality firms may sometimes find themselves with an empty queue due to the random service rate, on average, a random arrival faces a high probability of a long wait.

In many equilibria, at some queue lengths, agents play the same strategy, regardless of their signal. That is, agents with a good or bad signal join the queue with the same probability (either 0 or 1). Therefore, the equilibrium strategy at such queue lengths is independent of signal, or in other words displays herd-like behavior.

For $c > 0$, recall that $\bar{n}(c) - 1$ is the largest value of $n$ for which $\alpha^*(g, n) > 0$. Therefore, herd behavior is only interesting for $n \leq \bar{n}(c)$.

**Definition 2** An agent displays herd behavior at queue length $n \leq \bar{n}(c)$ if $\alpha^*(g, n) = \alpha^*(b, n)$.

\[^{4}\text{Threshold strategies of this sort are common in queuing models, since waiting costs are strictly increasing in queue length (see, for example, Naor, 1969, or Hassin and Haviv, 2002).}\]
This definition corresponds to Chamley’s (2004) definition of an agent herding when his action is independent of his signal. In our model, herding may occur at some queue lengths and not others (for example, Proposition 1 implies there is no herd behavior at queue length zero). Thus, herd behavior is necessarily local in our model. In a stationary equilibrium of our model, the empty queue is reached infinitely often as the number of agents grows toward infinity. This implies that we cannot have an informational cascade—a point of time beyond which all arriving agents herd. Starting at any queue length, random service departures will eventually reduce the queue to zero, at which point there is no herding.

3.1 Pure informational externality: \( c = 0 \).

To isolate the informational role of queues, suppose that \( c = 0 \). In this case, there are no congestion costs imposed by long queues, and the decision to join depends only on the information content in the queue length.

With no congestion costs, the equilibrium conditions simplify because \( \gamma(n) \) is independent of the queue length and is just \(-\frac{v_h}{v_f}\). This implies that an agent with a signal \( s \) always joins the queue if the likelihood ratio is above \(-\frac{v_h}{v_f}\). Further, the likelihood ratio is weakly decreasing in the queue length. Therefore, an agent with a bad signal may not join the queue if it is too short, but will join the queue if it is long enough. Depending on parameter values, there may be a case in which an agent with a bad signal is indifferent between joining and not over a non-singleton set of queue lengths, and chooses to join with certainty except at one specific queue length.

**Proposition 3** Suppose \( c = 0 \). In equilibrium:

(i) \( \alpha^*(g,0) > 0 \) and \( \alpha^*(g,n) = 1 \) for all \( n \geq 1 \).

(ii) there exists an \( n_b \in \{1, 2, 3\} \) such that \( \alpha^*(b,n) = 0 \) for \( n < n_b \) and \( \alpha^*(b,n_b) > 0 \).

(iii) \( \alpha^*(b,n) = 1 \) for all except at most one \( n \geq n_b \).

Thus, with \( c = 0 \), an agent with a bad signal plays a threshold strategy of the following form: he joins the queue only if the queue length is above some threshold. This is the converse of the usual queuing threshold strategy in which an agent enters only if the queue is sufficiently short. For \( n \geq n_b \), the equilibrium reflects herding, since agents with both good and bad signals join the queue with probability 1 (except for at most one value of \( n \) in this range).

The following example clarifies the tradeoff faced by agents when \( c = 0 \).

**Example 2** Let \( v_h = 1, v_f = -1, c = 0, p_0 = \frac{1}{2}, q = 0.7, \lambda = 1, \) and \( \mu = 1.3 \). For these

\(^5\)Chamley (2004), Definition 4.1.
parameters, there exists an equilibrium in which agents with good signals join the queue with probability 1 at all queue lengths, and agents with bad signals balk if the queue length is zero or one, but join with probability 1 if the queue length is two or higher. A comparison of the left and right hand sides of the best response condition (see Lemma 2) is shown in Figure 1 below.

\[ \text{Figure 1: Best responses in equilibrium when } c = 0. \text{ An agent with signal } s \text{ arriving at a queue length } n \text{ joins the queue only if } \psi(s, n) \leq \gamma(n). \]

Since \( c = 0 \), \( \gamma(n) \) is constant and equal to \( \frac{\psi_h}{\psi_r} = 1 \). In equilibrium, \( \psi_g(0) = 0.82 \) and \( \psi_b(0) = 4.46 \). Therefore, at a queue length of zero, only an agent with a good signal joins the queue. This then implies that \( r_h(0) = q \) and \( r_r(0) = 1 - q \), so that \( \psi_s(1) = \frac{1-q}{q} \psi_s(0) \) for each signal \( s \).

Now, at queue length 1, we still have \( \psi_b(1) > 1 \), so that an agent with a bad signal does not join the queue. This further implies that \( \psi_s(2) = \frac{1-q}{q} \psi_s(1) < 1 \) for each \( s \). At a queue length of 2, therefore, an agent with a bad signal joins with probability 1. Now we have \( r_h(2) = r_r(2) = 1 \), so that \( \psi_s(3) = \psi_s(2) \) for each \( s \). But then this implies again that both agents join a queue of length 2, and so on. Beyond \( n = 2 \), therefore, both \( \psi_g(n) \) and \( \psi_b(n) \) are constant in \( n \).
3.2 Information and congestion externalities: $c > 0$.

Suppose $c > 0$, so agents incur a positive disutility to waiting. There is now a congestion externality in addition to the informational externality. All else equal, an agent would prefer to face a shorter queue than a longer one. The tradeoff, of course, is that queue length is still informative about the quality of the firm.

With positive waiting costs, agents with bad signals no longer play threshold strategies. If there is any queue length at which an agent with a bad signal joins the queue with positive probability, a partial characterization of her strategy is as follows.\(^6\) There exists an upper threshold, $\tilde{n}_b > 0$, beyond which the agent does not join the queue, since waiting costs are too high. This threshold comes from the congestion costs implied by a queue. There also exists a lower threshold, $n_b > 0$, such that the agent will not join the queue unless there are at least $n_b$ already waiting. This threshold comes from the informational role of the queue. Whenever $n_b < \tilde{n}_b$, an agent with a bad signal joins at some queue length, but need not join for every queue length between the two thresholds. In particular, there can be “holes,” in their strategies. That is, an agent with a bad signal may not join the queue at some $n$, but, join at $n - 1$ and $n + 1$.

**Proposition 4** Suppose $c > 0$. Then,

(i) there exists an $\tilde{n}_g(c) \leq \tilde{n}(c)$ such that $\alpha^*(g, 0) > 0$ for all $n = \{0, 1, \cdots, \tilde{n}_g(c) - 1\}$ and $\alpha^*(g, \tilde{n}_g(c)) = 0$.

(ii) there exist parameter values for which agents with bad signals do not play threshold strategies.

The proof of part (ii) is by example.

**Example 3** Assume all other parameters are the same as in Example 2, and $c = 0.14$. There exists an equilibrium with the following strategies: an agent with a good signal joins the queue if and only if the queue length is less than 8, and an agent with a bad signal joins the queue at queue lengths 3, 5, and 6, but at no other lengths. There is, therefore, a “hole” in the strategy of the bad-signal agent when $n = 4$.

The intuition behind these strategies is represented in Figure 2. Recall that $\gamma(n)$ is inversely related to the cost of congestion at queue length $n$. In this example, $\gamma(n)$ declines non-linearly in $n$.

At queue lengths 3, 5, and 6, both the good and bad signal agents join thus $\psi_s(n)$ is flat, or beliefs are invariant to queue length, for each $s$. At a queue length of 3, $\psi_s(n)$ for

\(^6\)In Proposition 6 below, we demonstrate an equilibrium in which a bad-signal agent never joins.
Figure 2: Best responses in equilibrium when \( c = 0.14 \). An agent with signal \( s \) arriving at a queue length \( n \) joins the queue only if \( \psi(s, n) \leq \gamma(n) \).

Each \( s \) declines to a multiple \( \frac{1-q}{q} = \frac{3}{7} \) of its value at \( n = 2 \). Since an agent with a bad signal will join the queue at \( n = 3 \), \( \psi_b(n) \) is flat at this queue length. Meanwhile, \( \gamma(n) \) continues to decline, so that at queue length 4, \( \gamma(4) < \psi_b(4) \). But, only an agent with a good signal joins the queue at 4, so \( \psi_b(n) \) declines at \( n = 4 \), to the extent that \( \psi_b(5) < \gamma(5) \). Therefore, an agent with a bad signal joins a queue of length 5. A similar analysis indicates such an agent will join the queue at length 6, but not any higher length. Finally, at \( n = 9 \), no agent joins the queue, regardless of signal.

Thus, with positive waiting costs, a bad-signal agent’s strategy need not be a threshold strategy once queue length plays a dual role. On the one hand, a longer queue implies a greater likelihood the firm is high quality. However, a longer queue also imposes higher waiting costs on an agent who joins it. In equilibrium, the decision to join the queue trades off these two effects.
4 Social Learning

Since high and low quality firms generate different distributions over queue length, agents on average do learn about the quality of the good when they observe the number of agents waiting in the queue. However, learning is not perfect: The informativeness of the queue depends on its length. In a stationary equilibrium, every queue length from 0 to $\tilde{n}_g(c)$ has positive probability. Since agents do not know how many other agents arrived before them, their beliefs about quality do not converge as the number of arriving agents grows large. With positive probability agents choose an incorrect action—they either join when the good has low quality (analogous to a Type II error) or balk when it has high quality (analogous to a Type I error).

Given an equilibrium $\alpha^*$, let $\epsilon_h(\alpha^*)$ denote the probability that a randomly arrived agent makes an incorrect decision when the firm quality is high; that is, the probability that this agent refuses to join the queue (since the correct decision when quality is high is to join the queue). Similarly, let $\epsilon_\ell$ denote the probability that a randomly arrived agent makes an incorrect decision (i.e., joins the queue) when the firm quality is low.

Recall that $\tilde{n}(c)$ is the maximal queue length which can be observed in equilibrium. Exploiting the PASTA property,

$$
\epsilon_h(\alpha^*) = \sum_{n=0}^{\tilde{n}(c)} \pi_h(n; \alpha^*) \left[ q(1 - \alpha^*(g, n)) + (1 - q)(1 - \alpha^*(b, n)) \right] \quad (6)
$$

$$
\epsilon_\ell(\alpha^*) = \sum_{n=0}^{\tilde{n}(c)} \pi_\ell(n; \alpha^*) \left[ (1 - q)\alpha^*(g, n) + q\alpha^*(b, n) \right]. \quad (7)
$$

Thus, $\epsilon_h(\alpha^*) = \sum_{n=0}^{\tilde{n}(c)} \pi_h(n; \alpha^*)(1 - r_h(n; \alpha^*))$, and $\epsilon_\ell(\alpha^*) = \sum_{n=0}^{\tilde{n}(c)} \pi_\ell(n; \alpha^*) r_\ell(n; \alpha^*)$.

We find that there is an asymmetry in error rates whenever the equilibrium satisfies the following two properties: first, an agent with a good signal enters with probability 1 at every queue length up to $\tilde{n}_g(c) - 1$, and second, an agent with a bad signal joins the queue at some length with positive probability. Under these conditions, agents are more likely to err by acquiring the product when the quality is low rather than by not acquiring it when the quality is high. When $c = 0$, all equilibria satisfy these conditions, as demonstrated in Proposition 3. The conditions are also satisfied in Examples 3 and 4 (below), where $c$ is positive.
Proposition 5 Consider an equilibrium $\alpha^*$ in which (i) $\alpha^*(g,n) = 1$ for all $n \in \{0, \ldots, \bar{n}_g(c) - 1\}$, and (ii) there exists some $\hat{n}$ for which $\alpha^*(b, \hat{n}) > 0$. Then, a randomly arriving agent is more likely to enter the queue for a low quality good than fail to enter the queue for a high quality good, or $\epsilon_\ell(\alpha^*) > \epsilon_h(\alpha^*)$.

The asymmetry in learning with a high quality firm as opposed to a low quality firm is due to the one-sided nature of the servicing phenomenon. On being serviced, agents leave the queue. Therefore, a short queue may exist either because few people joined it, or because people who joined it were serviced relatively rapidly. An agent who arrives and sees a short queue cannot distinguish between these events, and is therefore more likely to err by joining a short queue. Recall that, on average, the queue for a low quality firm is shorter than that for a high quality firm. Therefore, agents are more likely to err (and join the queue) when the firm is low quality.

Every time an agent with a bad signal chooses not to join a queue, an increase in the queue length demonstrates that the joining agent had a good signal. All participants update accordingly. In some cases, when the congestion cost is positive, signals can be sufficiently informative that agents with bad signals do not join the queue at any length. At this point, agents enter solely based on their signals and there is no herding or social learning.

Proposition 6 Suppose $c > 0$ and $v_h < \frac{2c}{\mu}$. For $q$ sufficiently close to 1, there exists an equilibrium in which an agent with a good signal joins the queue if and only if $n = 0$, and an agent with a bad signal never joins. Therefore, no herding occurs.

It is immediate to observe that, for the equilibrium exhibited in Proposition 6, $\epsilon_h = \epsilon_\ell = 1 - q$. That is, the error rates are the same, regardless of the quality of the firm. Suppose we choose parameters for which a no-herding equilibrium exists. Consider the effect of keeping the other parameters fixed and reducing the congestion cost $c$. A reduction in $c$ has two effects: (i) the upper threshold for an agent with a good signal, $\bar{n}_g(c)$, increases and (ii) an agent with a bad signal joins the queue for some values of $n$. The former has no effect on the respective error rates, whereas the latter induces an asymmetry between the error rates for low and high quality firms. When $c$ falls to zero, the equilibrium displays herd-like behavior at a countably infinite number of queue lengths, and the asymmetry in error rates is at its highest.

5 Effect of changing the service rate

Suppose a firm could costlessly increase its service rate (for example, by adding extra capacity). Would it necessarily do so? In this section, we show that, over some parameter
range, the answer is no. Rather, a firm benefits from having customers waiting for the product, since a queue serves as a signal to new arrivals that the good has high quality.

Note that if firms could choose their service rates, the choice of \( \mu \) itself is a signal, and consumers update their beliefs on observing \( \mu \). The high and low quality firm must pool on \( \mu \)—if they do not, customers can immediately distinguish between the two firms, and the low quality firm will get no custom. If there are no costs to increasing the service rate, the low quality firm costlessly imitates the high quality firm, so that both firm types choose the same service rate. A similar argument applies if firms choose prices—they must pool on price. If prices are unchanged, maximizing profit is equivalent to maximizing the rate at which consumers join the queue for the product.\(^7\)

Consider the high quality firm. Since \( \epsilon_h(\alpha^*) \) denotes the rate at which a randomly arrived consumer does not join the queue, the joining rate for the firm is \( (1 - \epsilon_h(\alpha^*)) \) per arrival, or \( \lambda_h = \lambda(1 - \epsilon_h(\alpha^*)) \) per unit of time.

First, suppose \( c = 0 \). Here, a queue plays a purely informational role. The lower the capacity or service rate (\( \mu \)), the higher the endogenous arrival rate of agents. If firms are slow in servicing customers, the likelihood of shorter queues is decreased. As customers view short queues as negative signals, they are less likely to enter.

Therefore, intuition suggests that a high quality firm would wish to induce as long a queue as possible, or, in other words, choose a low service rate. However, comparative statics on \( \mu \) are problematic, as there may be multiple equilibria, and a change in \( \mu \) may result in a change in the nature of the equilibrium. Therefore, we focus on a particular equilibrium with \( \alpha^*(g,0) = 1 \) and \( \alpha^*(b,n) = 1 \) for all \( n \geq 1 \). Such an equilibrium exists when \( \frac{2p_0 - 1}{q + p_0} > \frac{\lambda}{\mu} \). We show that the joining rate for the high quality firm is indeed declining in \( \mu \).

**Proposition 7** Suppose that \( c = 0 \), \( v^h = -v^l \), and \( \frac{\lambda}{\mu} < \frac{2p_0 - 1}{q + p_0 - 1} \). Then, the following is an equilibrium strategy: \( \alpha^*(g,n) = 1 \) for all \( n \), \( \alpha^*(b,0) = 0 \), and \( \alpha^*(b,n) = 1 \) for all \( n \). Given the equilibrium \( \alpha^* \), the joining rate for the high quality firm, \( \lambda_h \), is decreasing in the service rate, \( \mu \).

\(^7\)If a firm could adjust its price to compensate for a changed service rate, both \( v_h \) and \( v_l \) (which are net of price) would change in our model, and the overall effects are unclear. A queue would still provide information about good quality, so a firm would not want to charge a price that made the average queue length very low.
6 Social Welfare

Next, consider overall consumer surplus. Recall that $W_j(\alpha^*; c)$ is the expected waiting time for a randomly arriving consumer when the firm’s quality is $v_j$, where $j \in \{h, \ell\}$. Further, if the firm has quality $v_j$, an agent who arrives and faces queue length $n$ joins with ex ante probability $r_j(n; \alpha^*) = q_{\alpha^*}(g, n) + (1 - q)\alpha^*(b, n)$. Using the PASTA property and the prior over firm quality, overall equilibrium consumer surplus per agent is defined as

$$\Omega(\alpha^*) = p_0 \left[ v_h \sum_{n=0}^{\overline{n}(c)} \pi_h(n; \alpha^*) r_h(n; \alpha^*) - cW_h(\alpha^*) \right] + (1 - p_0) \left[ v_\ell \sum_{n=0}^{\overline{n}(c)} \pi_\ell(n; \alpha^*) r_\ell(n; \alpha^*) - cW_\ell(\alpha^*) \right].$$

Consumer surplus per unit of time is then $\lambda \Omega(\alpha^*)$, since $\lambda$ agents are expected to arrive in each time interval of length one. Using the definitions of error rates $\epsilon_h$ and $\epsilon_\ell$, we can alternatively write consumer surplus per agent as

$$\Omega(\alpha^*) = p_0 [v_h(1 - \epsilon_h(\alpha^*)) - cW_h(\alpha^*)] + (1 - p_0) [v_\ell \epsilon_\ell(\alpha^*) - cW_\ell(\alpha^*)].$$

Now, suppose $c = 0$, and valuations are symmetric, so that $v_h = -v_\ell = v$. Then, consumer surplus per agent reduces to

$$\Omega(\alpha^*; 0) = v[p_0(1 - \epsilon_h(\alpha^*)) - (1 - p_0)\epsilon_\ell(\alpha^*)]$$
$$= v[p_0 - \{p_0\epsilon_\ell(\alpha^*) + (1 - p_0)(\epsilon_\ell(\alpha^*))\}].$$

The term in the curly brackets is the ex ante expected error rate. Whereas a high quality firm maximizing its own joining rate will minimize its own error rate, a social planner seeking to maximize consumer surplus will minimize the ex ante error rate. Hence, the firm and the planner will generally choose different service levels. Nevertheless, we show that, under the same conditions as used in Proposition 7, if the service rate is sufficiently high, overall consumer surplus declines as the service rate increases. There is thus a range of parameters, over which a social planner would also prefer a lower service rate.\footnote{Note that since $v_h, v_\ell$ are consumption utilities net of price, we assume that the planner charges the same price as the firm.}

Proposition 8 Suppose all conditions of Proposition 7 are satisfied, and consider the identified equilibrium $\alpha^*$. Given this equilibrium, there exists a threshold service rate $\hat{\mu}$ such that if the service rate is sufficiently high (i.e., $\mu \geq \hat{\mu}$), consumer surplus $\Omega$ is decreasing in the service rate $\mu$.\footnote{Note that since $v_h, v_\ell$ are consumption utilities net of price, we assume that the planner charges the same price as the firm.}
In general, the social planner will choose a different rate than that chosen by a high quality firm. The firm optimally chooses a distribution over histories that is relatively uninformative (low service rate) while a social planner would prefer a distribution that is more informative (higher service rate) to prevent agents incorrectly patronizing the low quality firm. Notice that the social planner does not necessarily choose the highest possible service rate, as an empty queue causes agents with bad signals to balk.

Do these results continue to hold when there are congestion costs, so that \( c > 0 \)? For \( c \) sufficiently close to zero, the intuition of Propositions 7 and 8 goes through. Although both the high quality firm and the social planner now need to consider waiting costs, over some parameter range, each prefers a lower service rate to a higher one. In traditional queueing models, congestion costs are the sole factor under consideration; all else equal, all parties would prefer a higher service rate. In our model, there is a tradeoff between the informational content of a queue and the extra costs it imposes, and sometimes the former dominate.

We show via a numeric example that, even when the waiting cost \( c \) is well above zero, both the high quality firm and the social planner may prefer a lower service rate to a higher one

**Example 4** The parameters are similar to those in Example 3. Consider \( \lambda = 1, v_h = -v_l = 1, p_0 = \frac{1}{2}, q = 0.7, \) and \( c = 0.14 \). We consider four values for the service rate: \( \mu = 1.34, 1.4, 1.72, 1.75 \).

In each case, there is a unique equilibrium, with agents playing a pure strategy. The joining rate per arrival for the high quality firm and the consumer surplus per arrival are shown in Table 1.

<table>
<thead>
<tr>
<th>Service rate ( \mu )</th>
<th>Agents’ strategies</th>
<th>Joining rate ( \lambda_h )</th>
<th>Consumer surplus ( \Omega )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.34</td>
<td>{0, 1, \ldots, 8}</td>
<td>0.737</td>
<td>0.095</td>
</tr>
<tr>
<td>1.4</td>
<td>{0, 1, \ldots, 8}</td>
<td>0.733</td>
<td>0.105</td>
</tr>
<tr>
<td>1.72</td>
<td>{0, 1, \ldots, 11}</td>
<td>0.725</td>
<td>0.136</td>
</tr>
<tr>
<td>1.75</td>
<td>{0, 1, \ldots, 11}</td>
<td>0.740</td>
<td>0.135</td>
</tr>
</tbody>
</table>

Table 1: Effects of changing service rate

Thus, (i) increasing the service rate from 1.34 to 1.4 reduces the joining rate for the high quality firm, but increases consumer surplus. The strategy played by agents in equilibrium remains the same. However, the increased service rate results in an increase in the steady state
frequency of observing no consumers in the queue. In this state, a bad signal agent does not join, reducing the joining rate for the firm.

\( ii \) increasing the service rate from 1.72 to 1.75 increases the joining rate for the high quality firm increase, but decreases consumer surplus decrease. In equilibrium, when the service rate is 1.75, a bad signal agent joins the queue at length 2. When the service rate is 1.72, she joins only at length 3. Since queue length 2 has higher probability, the effect is to reduce consumer surplus.

The example thus reinforces the notion that a queue communicates information about quality, and that this effect can dominate the effect of congestion costs. The presence of congestion costs makes the product less valuable for agents, so that they are less likely to purchase the product. Nevertheless, there are parameter values for which the high quality firm gains by reducing the service rate, and similarly there are parameter values for which consumer surplus too is higher when the service rate is lower.

Finally, it follows immediately from the example that, if a firm or planner could choose the service rate, in general they may choose different rates.

The example thus reinforces the notion that a queue communicates information about quality, and that this effect can dominate the effect of congestion costs. The presence of congestion costs makes the product less valuable for consumers, so that they are less likely to purchase the product. Nevertheless, there are parameter values for which the high quality firm gains by reducing the service rate, and similarly there are parameter values for which consumer surplus is higher when the service rate is lower.

Finally, it follows immediately from the example that, if a firm or planner could choose the service rate, in general they may choose different rates.

7 Conclusion

Much of the literature on social learning assumes that agents observe a truncated history (e.g., some random sample of history), rather than the seeing the entire history of the game. We argue that it is natural to consider agents who observe a truncated set of actions: they observe some agents who have chosen to purchase a product, but do not see how many took the opposite decision. In this framework, a queueing model provides both truncations. Arrival and service departures are exogenous and random, so agents arriving at different times will see different queue lengths. Further, agents who have joined the queue are all purchasers, so no decisions made by non-purchasers are seen.

We find that social learning in the queueing context depends on the queue length at which an agent arrives. If there are no waiting costs, for all queues beyond a certain length,
agents display herd behavior, so that queue length is uninformative. The queue length can become arbitrarily large. However in a stationary equilibrium, it must eventually drop down to a level at which only agents with good signals join the queue, so that social learning can occur.

With waiting costs, equilibrium behavior is richer. Agents with good signals continue to play threshold strategies. Agents with bad signals adopt strategies with “holes”—they may enter at queue lengths both below and above a particular level, but not at that level itself. Thus, at such holes, queues become informative and provide further information about the quality of the firm. Of course, the tradeoff is that longer queues are more costly for agents, due to the congestion externality.

One implication of the informational role of queues is that both a high quality firm and a social planner will sometimes prefer a low service rate to a high one, even when there are waiting costs for consumers. Thus, firms have an incentive to advertise a backlog of orders.
A Appendix: Proofs

Proof of Lemma 1

$\pi_j(n; \alpha)$ is the stationary probability of observing a queue of length $n$ when the firm has quality $j$ and agents play the strategy $\alpha$. This is the long run probability of a birth-death process. The rate at which agents join the queue when the queue length is $n$ is $\lambda r_j(n; \alpha)$. Once in the queue, customers leave at the rate $\mu$ (the death rate). Thus, given firm quality $j$ and agents’ strategy $\alpha$, the flow balance equations are

$$
\pi_j(n-1; \alpha) \lambda r_j(n-1; \alpha) + \pi_j(n+1; \alpha) \mu = \pi_j(n; \alpha) [\lambda r_j(n; \alpha) + \mu] \text{ for } n \geq 1
$$

$$
\pi_j(1; \alpha) \mu = \pi_j(0; \alpha) \lambda r_j(0; \alpha)
$$

Further,

$$
\sum_{n=0}^{\infty} \pi_j(n; \alpha) = 1
$$

Recursively solving this system of equations yields the expressions in the statement of the Lemma.

Proof of Lemma 2

An agent who faces state $(s, n)$ should join the queue with probability 1 if $u_s(n; \alpha) > 0$, or $\theta_s(n; \alpha) > \frac{-v_s + (n+1)c/\mu}{v_h - v_l}$. Since $\theta_s(n; \alpha) = \frac{p_s}{p_s \pi_h(n; \alpha) + (1-p_s) \pi_\ell(n; \alpha)}$, this condition reduces to

$$
\frac{p_s \pi_h(n; \alpha)}{p_s \pi_h(n; \alpha) + (1-p_s) \pi_\ell(n; \alpha)} > \frac{-v_\ell + (n+1)c/\mu}{v_h - v_\ell}. \tag{11}
$$

Cross-multiplying and rearranging terms leads to the condition stated in the Lemma.

The best responses when condition (11) holds as an equality, or with the inequality direction reversed, follow similarly.

Proof of Lemma 3

(i) Consider an agent who enters the system and sees a queue length $n$, when all other agents are playing the strategy $\alpha$. The probability that she will see queue length $n$ if the firm quality is $j$ is $\pi_j(n; \alpha)$, and is independent of her own signal (since it depends only on actions of previously arrived agents). From equation (2), we can write the posterior probability of the firm being high quality as

$$
\theta_s(n; \alpha) = \frac{\pi_h(n; \alpha)}{\pi_h(n; \alpha) + \frac{1-p_s}{p_s} \pi_\ell(n; \alpha)}.
$$
Since \( p_g > p_b \), it is immediate that \( \theta_g(n; \alpha) > \theta_b(n; \alpha) \), and hence \( u_g(n; \alpha) > u_b(n; \alpha) \).

Hence, an equilibrium strategy \( \alpha^* \) must satisfy \( \alpha^*(g, n) \geq \alpha^*(b, n) \). Now, since \( q > \frac{1}{2} \), from the definitions of \( r_h(n; \alpha) \) and \( r_\ell(n; \alpha) \), it follows that \( r_h(n; \alpha) \geq r_\ell(n; \alpha) \) for all \( n \).

(ii) Suppose that \( \alpha^*(g, n) = 0 \) for some \( n \). Since \( \alpha^*(b, n) \leq \alpha^*(g, n) \) (from part (i)), it follows that \( \alpha^*(b, n) = 0 \). Therefore, \( r_h(n; \alpha^*) = r_\ell(n; \alpha^*) = 0 \). Equation (4) now implies that \( \pi_h(\tilde{n}; \alpha^*) = \pi_\ell(\tilde{n}; \alpha^*) \) for all \( \tilde{n} \geq n + 1 \).

Proof of Proposition 1

(i) From Lemma 3 (i), \( r_h(n; \alpha^*) \geq r_\ell(n; \alpha^*) \) for all \( n \). Suppose now that \( r_h(n; \alpha^*) = r_\ell(n; \alpha^*) \) for all \( n \in N \). Then, Lemma 1 implies that \( \pi_h(n; \alpha^*) = \pi_\ell(n; \alpha^*) \) for all \( n \), including \( n = 0 \). Therefore, the queue length is uninformative, and an agent with signal \( s \) will join the queue only if

\[
\frac{p_s}{1 - p_s} \gamma(n) \geq 1.
\]

For \( n = 0 \), this condition reduces to

\[
\frac{p_s}{1 - p_s} \frac{v_h - c/\mu}{-v_l + c/\mu} < 1, \text{ or } p_s \geq \frac{v_l + c/\mu}{v_h - v_l}.
\]

From Assumption 1, part (ii), it follows that the last inequality holds strictly when \( s = g \), and is violated when \( s = b \). Hence, an agent with a good signal will join the queue when \( n = 0 \), and an agent with a bad signal will not. This contradicts the assumption that \( r_h(n; \alpha^*) = r_\ell(n; \alpha^*) \) for all \( n \). Therefore, it must be the case that there exists at least one \( \tilde{n} \) at which \( r_h(\tilde{n}; \alpha^*) > r_\ell(\tilde{n}; \alpha^*) \).

Now, note that \( \pi_j(0; \alpha^*) = \frac{1}{1 + \sum_{k=1}^{\infty} \left( \frac{\pi_h}{\pi_\ell} \right)^k r_j(k-1; \alpha^*)} \). Since \( r_h(n; \alpha^*) \geq r_\ell(n; \alpha^*) \) with strict inequality for at least one \( \tilde{n} \), it follows that \( \pi_j(0; \alpha^*) > \pi_h(0; \alpha^*) \).

(ii) Assumption 1, part (ii), implies that \( \frac{1-p_b}{p_b} < \gamma(0) \). Since \( \pi_\ell(0; \alpha^*) > \pi_h(0; \alpha^*) \), it follows a fortiori that \( \psi_b(0; \alpha^*) = \frac{1-p_b}{p_b} \frac{\pi_\ell(0; \alpha^*)}{\pi_h(0; \alpha^*)} < \gamma(0) \), so that \( \alpha^*(b, 0) = 0 \).

Suppose \( \alpha^*(g, 0) = 0 \). Since we have shown \( \alpha^*(b, 0) = 0 \) as well, a queue length of zero will be absorbing, and the system will never reach a higher queue length. It follows that \( \theta_s(0, \alpha^*) = p_b \). That is, the revised belief the firm is high quality, given queue length 0, is equal to the prior. But then \( u_g(0, \alpha^*) > 0 \) (by Assumption 1(ii)), which implies that \( \alpha^*(g, 0) = 1 \), which is a contradiction. Hence, \( \alpha^*(g, 0) > 0 \).

Proof of Proposition 2

(i) We first establish that for any \( n, \tilde{n} \) such that \( n < \tilde{n} \leq \tilde{n}(c) \),

\[
\frac{\pi_\ell(\tilde{n}; \alpha^*)}{\pi_h(\tilde{n}; \alpha^*)} \leq \frac{\pi_\ell(n; \alpha^*)}{\pi_h(n; \alpha^*)}.
\]
By definition, \( r_j(\hat{n}; \alpha) > 0 \) for all \( n \leq \hat{n}(c) \). Now, from the discussion following equation (4) for all \( n \leq \hat{n}(c) + 1 \), \( \pi_j(n + 1; \alpha) = \pi_j(n; \alpha) \frac{r_j(n; \alpha)}{r_j(n; \alpha)} \), thus

\[
\frac{\pi_\ell(n; \alpha)}{\pi_h(n; \alpha)} = \frac{\pi_\ell(0, \alpha)}{\pi_h(0, \alpha)} \prod_{k=0}^{n-1} \frac{r_\ell(k; \alpha)}{r_h(k; \alpha)}.
\]

(12)

From part (i) of lemma 3, \( \frac{r_\ell(n; \alpha)}{r_h(n; \alpha)} \leq 1 \) for \( n \leq \hat{n} - 1 \). Hence, for any \( n, \hat{n} \) such that \( n \leq \hat{n} \leq \hat{n} \), it follows that \( \frac{\pi_\ell(\hat{n}; \alpha)}{\pi_h(\hat{n}; \alpha)} \leq \frac{\pi_\ell(n; \alpha)}{\pi_h(n; \alpha)} \).

(ii) Now, since \( \pi_\ell(0, \alpha^*) > \pi_h(0, \alpha^*) \) and \( \frac{\pi_\ell(n; \alpha^*)}{\pi_h(n; \alpha^*)} \) is weakly declining in \( n \) for \( n \leq \hat{n}(c) \) and \( \pi_\ell(0, \alpha^*) > \pi_h(0, \alpha^*) \), it follows that the queue length distribution for firm \( h \) strictly first-order stochastically dominates the corresponding distribution for firm \( \ell \). Hence, the expected queue length for firm \( h \) is strictly higher than the expected queue length for firm \( \ell \). That is,

\[
\sum_{n=0}^{\hat{n}(c)} \pi_h(n; \alpha^*)n > \sum_{n=0}^{\hat{n}(c)} \pi_\ell(n; \alpha^*)n.
\]

(13)

(iii) From equation (5), it follows that, for all \( n \) and for \( j = h, \ell \),

\[
\pi_j(n + 1; \alpha^*) = \frac{\lambda}{\mu} \pi_j(n; \alpha^*)r_j(n; \alpha^*).
\]

(14)

Hence, \( \frac{\pi_j(n; \alpha^*)r_j(n; \alpha^*)}{\mu} = \frac{\pi_j(n+1; \alpha^*)}{\lambda} \), so that, for \( j = h, \ell \),

\[
W_j(\alpha^*; c) = \frac{1}{\lambda} \sum_{n=0}^{\hat{n}(c)} \pi_j(n + 1; \alpha^*)(n + 1).
\]

(15)

Now, \( \sum_{n=0}^{\hat{n}(c)} \pi_j(n + 1; \alpha^*)(n + 1) = \sum_{n=1}^{\hat{n}(c)} \pi_j(n; \alpha^*)n = \sum_{n=0}^{\hat{n}(c)} \pi_j(n; \alpha^*)n \). Thus, from equation (13), it follows that \( W_h(\alpha^*; c) > W_\ell(\alpha^*; c) \).

**Proof of Proposition 3**

(i) When \( c = 0 \), \( \gamma(n) = -\frac{\psi}{\psi_h} \), independent of \( n \). Further, from the proof of Proposition 1, \( \frac{\pi_\ell(n; \alpha^*)}{\pi_h(n; \alpha^*)} \) is weakly decreasing in \( n \) for each \( n \). Hence, \( \psi_j(n; \alpha^*) = \frac{1-p_\ell}{p_h} \frac{\pi_\ell(n; \alpha^*)}{\pi_h(n; \alpha^*)} \) is weakly decreasing in \( n \) for each \( n \) and each \( s = \{g, b\} \). Finally, it must be that \( \psi_g(n; \alpha^*) \leq -\frac{\psi}{\psi_h} \), else no agent joins the queue at 0, which would imply that \( \pi_\ell(0; \alpha^*) = \pi_h(0; \alpha^*) \). Since \( \alpha^*(g, 0) > 0 \) and \( \alpha^*(b, 0) = 0 \), we have \( \frac{r_\ell(0; \alpha^*)}{r_h(0; \alpha^*)} = \frac{1-q}{q} < 1 \). Therefore, \( \frac{\pi_\ell(0; \alpha^*)}{\pi_h(0; \alpha^*)} < \frac{p_\ell(0; \alpha^*)}{p_h(0; \alpha^*)} \). This now implies that \( \psi_g(1; \alpha) < -\frac{\psi}{\psi_h} \) so that \( \alpha^*(g, 1) = 1 \). In conjunction with the fact that \( \psi_g(n; \alpha) \) weakly declines in \( n \), we have that \( \alpha^*(g, n) = 1 \) for each \( n \geq 1 \).
We establish (ii) and (iii) simultaneously suppressing the dependence of $\pi_{\ell,\pi_h}$ on $\alpha^*$ for brevity. In the proof of Proposition 2 (ii), we show that, in equilibrium,

\begin{align}
\psi_g(0; \alpha^*) & = \frac{1 - p_g}{p_g} \frac{\pi_\ell(0)\pi_h(0)}{\pi_h(0)} \leq \gamma(0) \\
\psi_h(0; \alpha^*) & = \frac{1 - p_h}{p_h} \frac{\pi_\ell(0)\pi_h(0)}{\pi_h(0)} > \gamma(0).
\end{align}

When $c = 0$, $\gamma(n) = -\frac{v_h}{v_\ell}$ for all $n$. Further, $\frac{1 - p_0}{p_0} = \frac{1 - p_0}{p_0} \frac{q}{1-q}$, and $\frac{1 - p_h}{p_h} = \frac{1 - p_0}{p_0} \frac{1-q}{q}$.

Making these substitutions into the above equations yields

\begin{align}
\frac{1 - p_0}{p_0} \frac{q}{1-q} \frac{\pi_\ell(0)}{\pi_h(0)} & \leq -\frac{v_h}{v_\ell} \\
\frac{1 - p_0}{p_0} \frac{1-q}{q} \frac{\pi_\ell(0)}{\pi_h(0)} & > -\frac{v_h}{v_\ell}.
\end{align}

Since $\alpha^*(g, 0) > 0$ and $\alpha^*(b, 0) = 0$, we have $r_h(0) = q\alpha^*(g, 0)$ and $r_\ell(0) = (1-q)\alpha^*(g, 0)$. Hence, from equation (4),

\begin{align*}
\psi_b(1; \alpha^*) & = \frac{1 - p_b}{p_b} \frac{\pi_\ell(1)}{\pi_h(1)} = \frac{1 - p_0}{p_0} \frac{\pi_\ell(0)}{\pi_h(0)}.
\end{align*}

There are three cases to consider:

(a) Suppose $\frac{1 - p_0}{p_0} \frac{\pi_\ell(0)}{\pi_h(0)} < -\frac{v_h}{v_\ell}$, then $\alpha^*(b, 1) = 1$, and as $\frac{r_h(n)}{r_\ell(1)} \leq 1$ for all $n$, $\psi_g(n; \alpha^*) < -\frac{v_h}{v_\ell}$ for all $n \geq 1$. Thus, $\alpha^*(b, n) = 1$ for all $n > 1$. Here, $\overline{n}_b = 1$.

(b) Suppose $\frac{1 - p_0}{p_0} \frac{\pi_\ell(0)}{\pi_h(0)} = -\frac{v_h}{v_\ell}$. Then, $\alpha^*(b, 1) \in [0, 1]$. There are two sub-cases here:

(b1) If $\alpha^*(b, 1) < 1$, then $\frac{r_h(1)}{r_\ell(1)} < 1$ so that $\alpha^*(b, n) = 1$ for all $n > 1$. Hence, $\overline{n}_b = 1$.

(b2) If $\alpha(b, 1) = 1$, then $\frac{r_h(1)}{r_\ell(1)} = 1$, and hence $\psi_b(2; \alpha^*) = -\frac{v_h}{v_\ell}$; that is, at a queue length of 2, a bad signal agent is indifferent between entering and not. If $\alpha^*(b, 2) = 1$, then the same reasoning applies until the first queue length $n$ at which $\alpha(b, n) < 1$, at which point for all queues greater than $n$, the agent will enter with probability 1. Here, $\overline{n}_b = 1$.

(c) Suppose $\frac{1 - p_0}{p_0} \frac{\pi_\ell(0)}{\pi_h(0)} > -\frac{v_h}{v_\ell}$. Then, $\alpha^*(b, 1) = 0$. Since $\alpha^*(g, 1)$ is 1 from part (i) above, $r_h(1) = q$ and $r_\ell(1) = 1 - q$. Hence, $\psi_b(2; \alpha^*) = \frac{1 - p_b}{p_b} \frac{\pi_\ell(2)}{\pi_h(2)} = \frac{1 - p_0}{p_0} \frac{\pi_\ell(0)}{\pi_h(0)} \frac{q}{1-q}$. Note that this is exactly the LHS of equation 18. Hence, $\psi_b(2; \alpha^*) \leq -\frac{v_h}{v_\ell}$. Again, there are now three sub-cases to consider:

(c1) $\psi_g(2; \alpha^*) < -\frac{v_h}{v_\ell}$, so $\alpha^*(b, 2) = 1$. By the same logic as in (a) above, $\alpha^*(b, n) = 1$ for all $n \geq 3$ as well. Here, $\overline{n}_b = 2$.

(c2) $\psi_g(2; \alpha^*) = -\frac{v_h}{v_\ell}$, and $\alpha^*(b, 2) < 1$. Then, $r_h(2) > r_\ell(2)$, so $\psi_b(3; \alpha^*) = -\frac{v_h}{v_\ell}$, which implies that $\alpha^*(b, 3) = 1$. Further, since $r_\ell(n) \leq r_h(n)$ for all $n$, $\alpha^*(b, n) = 1$ for all $n \geq 3$. Here, $\overline{n}_b = 3$.

(c3) $\psi_g(2; \alpha^*) = -\frac{v_h}{v_\ell}$, and $\alpha^*(b, 2) = 1$. Then, $r_h(2) = r_\ell(2)$, so $\psi_b(3; \alpha^*) = -\frac{v_h}{v_\ell}$, which
implies that $\alpha^*(b, 3) \in [0, 1]$, so that a bad signal agent is again indifferent between joining and not. By the same logic as in (h2) above, if $\alpha^*(b, 3) < 1$, then $\alpha^*(b, n) = 1$ for all $n \geq 4$. If $\alpha^*(b, 3) = 1$, the process is repeated at a queue length of 4, and so on. Here, $\bar{n}_b = 3$. ■

Proof of Proposition 4

(i) Recall that when the queue is of length $\bar{n}(c)$, an agent will not join the queue, even if she knows the good has high quality. Further, $\bar{n}(c) \leq \frac{r_b}{c}$. Define $\bar{n}_g(c)$ as the smallest queue length at which $\alpha^*(\bar{n}_g(c)) = 0$. In equilibrium, an agent who faces queue length $n$ and has signal $s$ believes that there is a probability $\theta_s(n; \alpha^*)$ that the firm has high quality. Since $\theta_s(n; \alpha^*) \leq 1$, it is immediate that $\bar{n}_g(c) \leq \bar{n}(c)$. Further, it must be that $\alpha^*(g, n) > 0$ for all $n \in \{0, \ldots, \bar{n}_g(c) - 1\}$, else the queue length $\bar{n}_g(c)$ can never be reached.

(ii) The statement is proved via example. For the parameter values in Example 3, an agent with a bad signal does not follow a threshold strategy. ■

Proof of Proposition 5

Since $\alpha^*(g, n) = 1$ for all $n \leq \bar{n}_g(c) - 1$, it follows that $r_h(n) = q + (1 - q)\alpha^*(b, n)$ and $r_e(n) = 1 - q + q\alpha^*(b, n)$ for $n$ in this range. Further, $\pi_e(n; \alpha^*) = \pi_h(n; \alpha^*) = 0$ for $n > \bar{n}_g(c)$. From the definitions of $\epsilon_h, \epsilon_e$, we have

$$\epsilon_e(\alpha^*) - \epsilon_h(\alpha^*) = \sum_{n=0}^{\bar{n}(c)} [\pi_e(n; \alpha^*) (1 - q + q\alpha^*(b, n)) - \pi_h(n; \alpha^*) (1 - q - (1 - q)\alpha^*(b, n))]$$

$$= (1 - q) \sum_{n=0}^{\bar{n}(c)} [\pi_e(n; \alpha^*) - \pi_h(n; \alpha^*)] + \sum_{n=0}^{\bar{n}(c)} [\pi_e(n; \alpha^*) q + \pi_h(n; \alpha^*) (1 - q) \alpha^*(b, n)]$$

Since $\sum_{n=0}^{\bar{n}(c)} \pi_e(n; \alpha^*) = \sum_{n=0}^{\bar{n}_g(c)} \pi_h(n; \alpha^*) = 1$, the last equation reduces to

$$\epsilon_e(\alpha^*) - \epsilon_h(\alpha^*) = \sum_{n=0}^{\bar{n}(c)} [\pi_e(n; \alpha^*) q + \pi_h(n; \alpha^*) (1 - q) \alpha^*(b, n)] > 0,$$

since there exists an $\hat{n}$ for which $\alpha^*(b, \hat{n}) > 0$. ■

Proof of Proposition 6

Suppose $c > 0$ and $v < \frac{2c}{\mu}$. Consider the strategy $\alpha^*(g, 0) = 1$, $\alpha^*(g, n) = 0$ for all $n \geq 1$, and $\alpha^*(b, n) = 0$ for all $n$. We will show that if all other agents are playing $\alpha^*$, it is a best response for a randomly arrived agent to also play $\alpha^*$. Thus, $\alpha^*$ is an equilibrium strategy.

Suppose all previously arrived agents play $\alpha^*$. Then, a randomly arriving agent sees a queue length of either zero or one. Further, $r_h(0) = q$ and $r_e(0) = 1 - q$, with $r_h(n) = r_e(n) = 0$ for all $n \geq 1$. From Lemma 3, it follows that $\pi_e(0) = \frac{1}{1 + \lambda(1 - q)/\mu}$, and $\pi_h(0) = \frac{1}{1 + \lambda q}$. 

26
Thus, an agent with a good signal will join the queue at zero if
\[
\frac{1 - p_g}{p_g} \frac{\pi(0)}{\pi_h(0)} < \gamma(0), \quad \text{or,} \quad \frac{p_g}{1 - p_g} \frac{\pi_h(0)}{\pi(0)} \gamma(0) > 1. \tag{22}
\]
Similarly, an agent with a bad signal will not join the queue at zero if
\[
\frac{1 - p_b}{p_b} \frac{\pi(0)}{\pi_h(0)} > \gamma(0), \quad \text{or,} \quad \frac{p_b}{1 - p_b} \frac{\pi_h(0)}{\pi(0)} \gamma(0) < 1. \tag{23}
\]
Since \( \frac{p_g}{1 - p_g} = \frac{p_0}{1 - p_0} \frac{q}{1 - q} \) and \( \frac{p_b}{1 - p_b} = \frac{p_0}{1 - p_0} \frac{1 - q}{q} \), the conditions for a good signal agent to join and a bad signal agent to not join respectively reduce to:
\[
\frac{\gamma(0)}{p_0} \frac{1 + \lambda(1 - q)/\mu}{1 + \lambda q/\mu} \frac{q}{1 - q} > 1 \tag{24}
\]
\[
\gamma(0) \frac{p_0}{1 - p_0} \frac{1 + \lambda(1 - q)/\mu}{1 + \lambda q/\mu} \frac{1 - q}{q} < 1. \tag{25}
\]
Since \( \gamma(0) \) is bounded and positive, given any parameter values there exists a \( q \) sufficiently close to one such that both conditions (24 and (25) are satisfied. Hence, if all other agents play \( \alpha_0 \), a randomly arrived agent should also play \( \alpha^*(g, 0) = 1 \) and \( \alpha^*(b, 0) = 0 \).

Now, note that \( \gamma(n) = -\frac{\psi_h(n) - (n+1)c/\mu}{\psi_h(n)} < 0 \) when \( n \geq 1 \). Hence, when \( n \geq 1 \), we trivially have \( \gamma(n) < \psi_s(n) \), so that an agent’s best response is to not join the queue.

Therefore, if all other agents are playing \( \alpha^* \), it is a best response for a randomly arriving agent to also play \( \alpha^* \). Hence, \( \alpha^* \) constitutes an equilibrium.

It is immediate that \( \alpha^* \) displays no herding, since \( \bar{n}(c) = 0 \) and \( \alpha^*(g, 0) \neq \alpha^*(b, 0) \).

**Proof of Proposition 7**

First, we show that \( \alpha^* \) as defined in the statement of the Proposition is indeed an equilibrium. Suppose all other players are playing \( \alpha^* \), and let \( \alpha \) denote the best response of a randomly arriving player who sees \( n \) agents in the queue. Given \( \alpha^* \), we have \( r_h(0) = q \), \( r_\ell(0) = q \), and \( r_h(n) = r_\ell(n) = 1 \) for all \( n \geq 1 \). Hence,
\[
\pi_h(0) = \frac{1}{1 + q \sum_{n=1}^{\infty} \left( \frac{\lambda}{\mu} \right)^n} = \frac{1}{1 + \frac{\lambda}{\mu} \sum_{n=1}^{\infty} \left( \frac{\lambda}{\mu} \right)^n} = \frac{1 - \lambda/\mu}{1 - (1 - q) \lambda/\mu}, \tag{26}
\]
where the second equality uses the fact that \( \frac{\lambda}{\mu} < 1 \) when \( c = 0 \).

Similarly,
\[
\pi_\ell(0) = \frac{1}{1 + q \sum_{n=1}^{\infty} \left( \frac{\lambda}{\mu} \right)^n} = \frac{1 - \lambda/\mu}{1 - (1 - q) \lambda/\mu}. \tag{27}
\]
Hence, \( \frac{\pi_\ell(0)}{\pi_h(0)} = \frac{1 - (1 - q) \lambda/\mu}{1 - q \lambda/\mu} \).

27
Now, suppose \( \frac{\lambda}{\mu} < \frac{2p_0 - 1}{q + p_0 - 1} \). Suppose further that \( p_0 < \frac{1}{2} \), since \( q > \frac{1}{2} \). Hence, \( \frac{2p_0 - 1}{q + p_0 - 1} > 1 \), so it cannot be (given Assumption 1 (i)) be \( \frac{\lambda}{\mu} < \frac{2p_0 - 1}{q + p_0 - 1} \). Hence, it must be that \( q + p_0 - 1 > 0 \).

Going back to the condition \( \frac{\lambda}{\mu} < \frac{2p_0 - 1}{q + p_0 - 1} \) and multiplying both sides by \( q + p_0 - 1 \), we have

\[
\frac{\lambda}{\mu} (q + p_0 - 1) < 2p_0 - 1
\]

\[
\frac{\lambda}{\mu} [p_0q - (1 - p_0)(1 - q)] < p_0 - (1 - p_0)
\]

\[
\frac{1 - p_0}{p_0} \frac{1 - (1 - q)\lambda/\mu}{1 - q\lambda/\mu} < 1.
\]  

(28)

Now, note that \( \frac{1 - (1 - q)\lambda/\mu}{1 - q\lambda/\mu} = \frac{\pi_\ell(0)}{\pi_h(0)} \), \( \frac{v_h}{v_\ell} = 1 \), \( \frac{r_\ell(0)}{r_h(0)} = \frac{1 - q}{q} \), and \( \frac{1 - p_0}{p_0} = \frac{(1 - p_0)q}{p_0(1 - q)} \). Substituting these into (28), we have

\[
\frac{1 - p_0}{p_0} \frac{\pi_\ell(0)}{\pi_h(0)} \frac{r_\ell(0)}{r_h(0)} < -\frac{v_h}{v_\ell}.
\]  

(29)

Hence, \( \alpha(b, 1) = 1 \) is a best response to \( \alpha^* \). Further, implies that \( r_h(1) = r_\ell(1) = 1 \), so that \( \alpha(b, 2) = 1 \), and so on for all \( n \geq 1 \). Hence, \( \alpha(g, n) = 1 \) for all \( n \geq 1 \).

Consider \( \alpha(g, 0) \). Since \( q > \frac{1}{2} \), it follows from equation (28) that

\[
\frac{1 - q}{q} \frac{1 - p_0}{p_0} \frac{1 - (1 - q)\lambda/\mu}{1 - q\lambda/\mu} < 1
\]

\[
\frac{1 - p_0}{p_0} \frac{\pi_\ell(0)}{\pi_h(0)} \frac{r_\ell(0)}{r_h(0)} < -\frac{v_h}{v_\ell},
\]  

(30)

so that \( \alpha(g, 0) = 1 \).

Hence, \( \alpha^* \) is a best response to itself, and is therefore an equilibrium strategy.

Now, \( \pi_j(n + 1) = \pi_j(n)\frac{\lambda}{\mu}r_j(n) \), for each \( n = 0, 1, \ldots \) and \( j = \ell, h \). Hence, \( \pi_j(n)r_j(n) = \pi_j(n + 1)\frac{\lambda}{\mu} \). Thus,

\[
\lambda_h = \frac{\mu}{\lambda} \sum_{n=0}^{\infty} \pi_h(n + 1) = \frac{\mu}{\lambda} (1 - \pi_h(0))
\]

\[
= \frac{q}{1 - (1 - q)\lambda/\mu},
\]

which is clearly decreasing in \( \mu \).

\[\Box\]

**Proof of Proposition 8**

As shown in Proposition 7, in the identified equilibrium \( \alpha^* \), we have \( \lambda_h = \frac{q}{1 - (1 - q)\lambda/\mu} \).

Using the same technique, we can show that \( \lambda_\ell(\alpha^*) = \frac{1 - q}{1 - q_\ell/\mu} \).
Set $c = 0$ and $v_h = -v_\ell = v$. Then, consumer surplus $\Omega$ reduces to

$$
\Omega = v[p_0 \lambda_h + (1 - p_0) \lambda_\ell].
$$

Using the expressions for $\lambda_h, \lambda_\ell$ above,

$$
\frac{d\Omega}{d\mu} = v \left[ p_0 \frac{d\lambda_h}{d\mu} - (1 - p_0) \frac{d\lambda_\ell}{d\mu} \right]. 
$$

Hence, $\frac{d\Omega}{d\mu} < 0$ if and only if

$$
\frac{1 - p_0}{(1 - (1 - q) \lambda/\mu)^2} < \frac{p_0}{(1 - q \lambda/\mu)^2}.
$$

As $\mu \to \infty$, the latter inequality reduces in the limit to $1 - p_0 < p_0$, or $p_0 > \frac{1}{2}$.

We show that the conditions of Proposition 7 imply that $p_0 > \frac{1}{2}$. Consider the condition

$$
\frac{\lambda}{\mu} < \frac{2p_0 - 1}{q + p_0 - 1}.
$$

As shown in the proof of Proposition 7, if this condition is satisfied, then $q + p_0 - 1 > 0$. Since $\lambda, \mu$ are positive by definition, it must be that $2p_0 - 1 > 0$, or $p_0 > \frac{1}{2}$.

Hence, as $\mu \to \infty$, in the limit $\frac{d\Omega}{d\mu} < 0$. Since $\Omega$ is continuous in $\mu$, it follows there is a $\hat{\mu}$ large enough such that for all $\mu \geq \hat{\mu}$, $\frac{d\Omega}{d\mu} < 0$ and the consumer surplus decreases in the service rate.
References


