We study how a high quality service firm selects a service rate differently than a low quality service firm when the firm cannot communicate its service value or service rate to its customer base. As a result, potential customers may take the queue length upon arrival into account when assessing the service value before joining the queue for service. We show that customer queue joining behavior may not be of the threshold type, which is a typical equilibrium structure under observable service value and rate. Furthermore, we find differentiating equilibria in which the high quality service firm selects a slower service rate than the low quality service firm, even if the cost of speeding up is the same for both firms. We also find that both firm’s profit rates under such a differentiating equilibrium may be higher than in equilibria in which both firms select the same service rate. Our research thus provides a rational explanation for often-quoted suspicions that some firms deliberately slow down service to signal quality.

Key words: Strategic consumer behavior, queueing theory

1. Introduction

Due to the highly experiential nature of many services, it is often difficult for service firms to communicate precisely and credibly the value of the service they provide to customers. In such environments, it is natural that potential customers seek to update their prior about the service value with additional information from sources such as on-line reviews, consumer reports, advice from friends or colleagues, etc. One good indicator of the service value is the level of congestion at a service provider, which, in many service environments is the length of the queue a customer observes on her arrival. The queue length indicates how many other customers have found the service value high enough to overcome the inconvenience associated with waiting. Hence, when information about the value of the service is missing, the decision making of consumers is augmented by queue length information.
The service rate is a key determinant of the queue length formation and evolution. It is thus an essential piece of information that helps customers compare the service value and waiting cost inferred from the queue length. Nevertheless, the service rate of a process often cannot be easily observed by potential customers. An arriving potential customer would be able to observe the queue length with a cursory glance. However, to estimate the service capacity\textsuperscript{1}, she would have to spend time inspecting the service process, since it may be comprised of investments in technology, labor and organizational structure. For an example, a customer can infer the congestion at a restaurant by looking into the restaurant through the window, but she cannot easily obtain information related to the service rate (such as the number of employees at the restaurant, the size of the backroom operations etc.)\textsuperscript{2}. Such firm-specific information, unlike queue length, is not a natural outcome of the service process. Nevertheless, the extant research in service operations has often assumed service capacities to be observable or common knowledge, even when the queue lengths are unobservable (see Hassin and Haviv, 2003). In this paper, we relax this key assumption and allow the firm’s service rate to remain unobservable.

Knowing the paucity of information in the market, the service firm can strategically select service rates to influence customers’ queue joining behavior. In response, rational customers may factor in the firm’s strategies to build their inferences of service value from queue lengths. Thus, the queue joining strategy of rational customers and the firm’s service strategy are inherently intertwined problems.

In this paper, we formally address these intertwined problems. In our model, a firm provides a service that can be of either high or low value to an agent that is a potential customer. The ‘high quality firm’ provides higher value to the agents than the ‘low quality firm.’ Risk-neutral agents arrive stochastically at a market, and are served in a first-come, first-served queue. The arriving agent sees the number of people currently in the queue at the server. There is a cost associated

\textsuperscript{1} We use the term ‘service capacity’ synonymously with ‘service rate’.

\textsuperscript{2} Sometimes service provider announces the expected waiting time. However, such communication need not be credible (See Allon et al. (2007)). Instead, we focus on the service rate selection as the main decision variable of the service firm. In Allon et al. (2007), the service rates remain observable.
with waiting incurred by agents that have decided to join the queue. Each agent’s decision to join the queue or to balk is based on her inferences of service value and waiting costs.

The firm is perfectly aware of the value of its service, and controls the service process through its strategic service rate decisions. The firm can select either a fast or a slow service process – a selection that remains unobservable to the agents. For simplicity, we assume that the time to service an agent under both processes is exponentially distributed, and independent for each agent. We assume that there is a cost (per unit of time) of selecting the fast service process, i.e. there is a cost incurred for “speeding-up”. In addition, this cost may depend on the service value offered by the firm.

Our paper adds to the stream of research on queueing behavior. Equilibrium joining strategies in queues have been examined beginning with seminal work by Naor (1969), see Hassin and Haviv, 2003, for an excellent overview. When there are waiting costs, customers follow a threshold strategy: they join the queue as long as it is not too long. Beyond the threshold, congestion effects dominate any value from service (as in Edelson and Hildebrand 1975, cf Hassin and Haviv (2003)).

The economics of strategic delays and the transmission of delay information in queueing settings have garnered significant interest in recent years. A number of papers (Afeche (2005), Allon et al (2007), Guo and Hassin (2009), Guo and Zipkin (2007), Haviv (2009), Veeraraghavan and Debo (2008)) examine delay information in queueing settings. In this literature stream, the service value and rate are common knowledge (even in those cases when the queue lengths remain unobservable). To the best of our knowledge, no research addresses endogenous service rate selection of firms under unobservable service rates and unknown service values.

Ata and Shneorson (2006) study service rate decisions to maximize social welfare maximization in a queueing system, and state that an interesting open problem would be to analyze “sustainable equilibria given that customers have their own beliefs about the system state” under profit.

3 Speeding-up can occur through deployment of technology (e.g. automation of some service steps), or through hiring of extra personnel for back-end operations (for e.g. more cooks may be hired to work at a kitchen).

4 Services that provide high value may require more care and attention, and may have higher costs. As such, it may be more expensive for a high quality firm than for a low quality firm to increase the speed of the service, while maintaining the same quality. For example, higher skilled labor may be scarce and more expensive to hire. Automation equipments for high value services could be more expensive because of the need to have tighter tolerances.
maximization objectives. Our focus is on services where the service rate information is absent, and its subsequent effect on firms’ profits. Specifically, our paper aims to clarify the long-held question of whether strategic delays help signal service quality and improve firms’s profits. The answer is affirmative. Our paper makes the following contributions to the literature:

1. We relax two key assumptions in the classical queueing games models: observability of service rates and service values. Our model can be considered as expanding on the queueing games literature by introducing uncertainty pertaining to both the value of the service and service rate. We argue that while it is often possible to see the exact queue lengths before joining the service, the exact service rate remains unknown.

2. We show that queue joining behavior of rational customers under unobservable service value and service rates may result in a “sputtering” equilibrium, where the congestion (queue length) at the firm generally resides within a balking threshold, but may probabilistically break through to a strictly higher threshold value, depending on the choice of the customer who mixes between joining and balking at that particular state. This “sputtering” equilibrium fundamentally expands on the threshold equilibria seen in queues with decision-making customers.

3. We show how firms can strategically select their service rates to signal their service value to a market when customers do not observe service rates or value. When making decisions about whether to speed-up or slow down, the service provider needs to weigh the operational benefits accrued from faster service (the expected waiting time is lower, so more customers may join), against signaling benefits from slower service (congestion signals high service value, so more customers may join). In particular, firms might slow down the service rate to signal quality. The slowing down of the service rate by the high quality service firm is reminiscent of the finding by Afeche (2005), who shows that firms may want to artificially insert delays in a market with heterogeneous time-sensitive agents. In our model, all agents are homogenous – nevertheless, the uncertainty in information creates an incentive for the firm to slow down.

4. We show that even when the cost of speeding up does not depend on the firm’s service value, the unobservability of the service rate may lead to an equilibrium in which a firm randomizes
between offering slow and fast service rates with a high quality firm’s expected service rate being lower than the low quality firm’s expected service rate. Interestingly, this randomization of the service rate leads to customer learning from queues: customers may positively update the service value when seeing longer queues. The firm’s profit rate in such equilibria are higher than the profits in pure strategy service rate equilibria. In other words, we demonstrate that the unobservability of the service rates triggers customer learning from queues. Both the high and low quality firm profit from this learning. Therefore, “garbling” of service rate information leads to higher profits for firms and more learning for customers.

5. When the cost of speeding-up increases with the value offered by the service, firms of different quality differentiate themselves by choosing different service strategies. Thus, customer learning is enhanced.

6. Finally, due to difficulty in signaling speed or quality, it is possible that the profits of a high quality firm are lower than that of a low quality firm in the pure-strategy equilibrium, especially when the low quality can mimic the strategies of the high quality firm.

The rest of the paper proceeds as follows. In Section 2, we provide a formal description of our model. In Section 3, we characterize agents’ queue joining strategy, given the service strategy of the firm. In Section 4, we analyze the firms’ decision strategies by employing the agents’ strategy that we analyzed in the previous section. We finally summarize by discussing our insights, and pointing out future directions in Section 5.

2. Model

We consider a firm selling an experience good. The firm can be one of two types, \( \omega \in \{ h, \ell \} \). The utility an customer obtains from purchasing and consuming a good from a firm with type \( \omega \) is \( v_\omega \), with \( v_h \geq v_\ell \). This utility is net of the good’s price, which is not explicitly modeled.

Agents are risk-neutral, and arrive sequentially at the market according to a Poisson process with parameter \( \Lambda \). It takes time to service (i.e., provide the good to) each agent. Thus, if the arriving agents cannot be immediately served, they wait, and hence, form a queue. The queue is
served on a first-come, first-served basis. All agents suffer a disutility (in the form of waiting cost) of \( c > 0 \) per unit time while waiting to obtain the good. The agent’s prior about the firm’s type is \( \Pr(\omega = h) = p \). After agents observe the queue length upon their arrival, \( n \), they decide whether to join the queue or balk.

The firm can select one of two service processes with exponentially distributed service times: \( 1/\mu \) (1/\( \overline{\mu} \)) is the mean service time of the slow (fast) process (i.e. \( \mu > \overline{\mu} \)). Without any loss of generality, we normalize the cost of the slow process to \( K_h(\mu) = K_f(\mu) = 0 \). We denote \( K_h(\overline{\mu}) \doteq k_h \) and \( K_f(\overline{\mu}) \doteq k_f \). Every agent that joins the queue generates a revenue of \( r \) for the firm. The firm maximizes the long run average profit rate. The agents maximize their expected net utility (i.e. utility of obtaining the good minus the expected waiting costs), conditional on their available information. The model parameters are \( (\Lambda, p, v_\ell, v_h, r, k_h, k_f, \overline{\mu}, \mu) \).

**The Players’ Strategies and Beliefs:** All agents decide the probability of joining the service after observing the queue length \( n \), denoted by \( \alpha(n) \). The agent’s updated belief that the type of the firm is high after observing a queue of length \( n \) is denoted by \( \gamma(n) \).

The firm decides the probability of selecting the fast, \( \overline{\mu} \), (or slow, \( \mu \)) service rate after observing the type \( \omega \), is denoted by \( \beta(\omega) \) \( (1 - \beta(\omega)) \). When high service rate \( \overline{\mu} \) is chosen by a firm, we have \( \beta(\omega) = 1 \). In short, \( \alpha = (\alpha(0), \alpha(1), \alpha(2), ...), \ \beta = (\beta(h), \beta(\ell)) \) and \( \gamma = (\gamma(0), \gamma(1), \gamma(2), ...) \) denote the agent’s joining strategy, the firm’s service rate selection and the agent’s updated belief respectively. We need to determine the equilibrium strategies of the players (the firm and all agents): \( \alpha^*, \beta^*, \) and \( \gamma^* \).

**The Equilibrium Conditions:** We now develop the equilibrium concept and the conditions for the equilibrium, to determine the strategies of the firm and the agents.

Consider a randomly arriving agent. Suppose all other agents are joining according to some strategy \( \alpha \). The agent’s expected utility of joining the queue after observing the queue length \( n \), is a function of the queue length, its belief about the firm’s type, \( \gamma \), and the firm’s service rate strategy, \( \beta \), denoted by: \( u(n, \gamma, \beta) \). In a symmetric equilibrium, the agent’s expected payoff in state \( n \) from joining with probability \( \alpha \) is \( \alpha \times u(n, \gamma, \mu) \). Finally, a firm’s expected revenue rate
Debo and Veeraraghavan: Unknown Service Quality and Speed

(per unit of time) when all players play $\alpha$ and the firm’s service rate is $\mu$ is denoted by $R(\alpha, \mu)$.

Now, we describe the equilibrium conditions:

**DEFINITION 1.** The strategies $\alpha^*, \beta^*$ and beliefs $\gamma^*$ form a Markov Perfect Bayesian Equilibrium (Maskin and Tirole 2001) if:

(i)-agents are rational: for each $n \in \mathbb{N},$

\[ \alpha^* (n) \in \arg \max_{\alpha' \in [0,1]} \alpha' \times u(n, \gamma^*, \beta^*). \]  

(ii)-beliefs are consistent: The belief $\gamma^* (n)$ satisfies Bayes’ rule on all queue lengths that are reached with strictly positive probability on the long run under strategies $\alpha^*$ and $\beta^*$.

(iii)-firms are rational:

\[ \beta^*(\omega) \in \arg \max_{\beta \in [0,1]} \beta(R(\alpha^*, \mu) - k_\omega) + (1 - \beta)R(\alpha^*, \mu) \text{ and } \omega \in \{\ell, h\}. \]

The conditions (i) and (iii) of Definition 1 are referred to as the *rationality* conditions for the agents and firms respectively. The condition (ii) of Definition 1 is referred to as the *consistency* condition of the beliefs.

**Discussion of the Model:** Notice that a key innovation in our model with respect to the existing literature is Equation (2). It captures our observation that the service rate is typically not common knowledge. Instead, in our model, it is endogenously determined.

When $k_h = k_\ell$ and $v_h > v_\ell$, the cost of speeding up is common knowledge. This case allows us to study the joint impact of service value uncertainty and unobservable service rates.

If both $v_h > v_\ell$ and $k_h > k_\ell$, in addition to the unobservable service rate, agents are uncertain about the quality and the cost of speeding up, except that they believe it is expensive for high quality firms to speed up and maintain the same high quality service.

---

5 When a queue length is not reached with positive probability, the belief is irrelevant, as well as the action at that queue length.
3. Analysis of Agent Joining Behavior for a Given Service Rate Strategy

In this section, we introduce the notion of an ‘agent joining equilibrium,’ which satisfies conditions (i) and (ii) of Definition 1 for a given service rate strategy characterized by $\beta$. We indicate the agent joining equilibrium quantities, $\bullet$, by means of a hat, $\hat{\bullet}$. In the next section, we then also impose condition (iii) of Definition 1 and characterize the equilibrium service rate selection of the firm. For convenience we focus on pure strategy equilibria in service rates and use $\mu_{ff} = (\mu, \mu)$, $\mu_{ss} = (\mu, \mu)$, $\mu_{sf} = (\mu, \mu)$, and $\mu_{fs} = (\mu, \mu)$ as a shorthand notation for pure strategy service rate selections, $(1, 1)$, $(0, 0)$, $(0, 1)$ and $(1, 0)$ respectively. For any strategy $\mu$, with $\sigma \in \{ss, ff, sf, fs\}$, the first letter consistently refers to the speed of the high quality firm (slow or fast), while the second letter refers to the speed of the low quality firm (slow or fast). Thus, as an example, $\mu_{sf} = (\mu, \mu)$ implies the condition when the high quality firm would choose slow service speed, and the low quality firm chooses fast service speed. Thus the agent’s utility can be written as:

$$u(n, \gamma, \beta) = \frac{\gamma(n)}{1-\gamma(n)} v_h - c(n+1) \frac{1}{\mu_h} v_f - \frac{c(n+1)}{\mu_h} \left( v_f - c(n+1) \frac{1}{\mu_f} \right).$$

(3)

With Equation (3), we can rewrite condition (i) of Definition 1 for agents in terms of the ratio of the updated probability that the firm is of high quality when an agent observes queue length $n$ over the updated probability that the firm is of low quality when the agent observes queue length $n$:

$$u(n, \gamma, \beta) > 0 \iff \frac{\gamma(n)}{1-\gamma(n)} > \frac{-v_f + c(n+1) \frac{1}{\mu_f}}{v_h - c(n+1) \frac{1}{\mu_h}}.$$

(4)

Now, instead of the updated belief, $\gamma(n)$, it will be convenient to introduce the likelihood ratio, $l(n)$, satisfying: $\gamma(n) / (1 - \gamma(n)) = l_0 \times l(n)$, where $l_0 = p / (1 - p)$. In other words, we decompose $\gamma(n)$ into the agent’s prior likelihood (that the firm is of high quality), $l_0$, and the likelihood ratio of observing $n$ agents upon arrival, $l(n)$. Then, we can rewrite Equation (4):

$$u(n, \gamma, \beta) > 0 \iff l(n) > L(n, \mu),$$

where $L(n, \mu) = \frac{1}{l_0} \frac{-v_f + c(n+1) \frac{1}{\mu_f}}{v_h - c(n+1) \frac{1}{\mu_h}}.$

(5)
$L$ can be interpreted as the \textit{minimum likelihood ratio} that is required to make an agent join a queue of length $n$ when the firm’s service strategy is $\mu$. The rationality condition (i) of Definition 1 determines $\alpha(n)$ for any $n$, which is 1 (0) when $l(n) > L(n, \mu)$ ($l(n) < L(n, \mu)$).

Now, we discuss condition (ii) of Definition 1. This condition needs to be imposed on $l(n)$. To that end, let $\pi(n, \alpha, \mu)$ be the long run probability that $n$ agents are in the system when the agent strategy is $\alpha$ and the firm’s service rate is $\mu$. For a given service rate and agent joining strategy, the stochastic process that describes the queue length is a Birth-Death process. The following Lemma characterizes the long run probability distribution of the queue length (all proofs are in the Appendix).

\textbf{Lemma 1.} Suppose all agents follow the strategy profile $\alpha$. Then, for both $\omega \in \{h, \ell\}$, the stationary probability of observing a queue of length $n$ is $\pi(n, \alpha, \mu)$

\[
\pi(n, \alpha, \mu) = \pi(0, \alpha, \mu) \prod_{j=0}^{n-1} \frac{\alpha(j) \Lambda}{\mu} \text{ where } \pi(0, \alpha, \mu) = \left[ 1 + \sum_{n=1}^{\infty} \prod_{j=0}^{n-1} \frac{\alpha(j) \Lambda}{\mu} \right]^{-1}.
\]

With the PASTA property (Wolff, 1982), $\pi(n, \alpha, \mu)$, is also the probability that a randomly arriving agent observes $n$ agents in the system. The posterior probability that the firm’s quality is high after observing a queue length of $n$ is $p \pi(n, \alpha, \mu)_{h}$ for all queue lengths that are reached with positive probability in equilibrium\footnote{$l(n)$ is indeterminate whenever $\pi(n, \alpha, \mu)_{h}$ and $\pi(n, \alpha, \mu)_{\ell}$ are both zero.}. With Lemma 1, notice that $\pi(n, \alpha, \mu)_{h}$ becomes equal to $\pi(0, \alpha, \mu)_{h} \left( \frac{\mu_{\ell}}{\mu_{h}} \right)^{n}$. As the first factor is equal to $l(0)$, the consistency condition Definition 1(ii) becomes:

\[
l(0) = \frac{\pi(0, \alpha, \mu)_{h}}{\pi(0, \alpha, \mu)_{\ell}} \text{ and } l(n) = l(0) \left( \frac{\mu_{\ell}}{\mu_{h}} \right)^{n} \text{ for } n \geq 1. \tag{6}
\]
because the likelihood ratio at the empty queue, \( l(0) \), along with Equation (6) immediately determines \( l(n) \) for any \( n \geq 1 \). With Equation (5), the rational actions can be immediately determined by comparing \( l(n) \) with \( L(n, \mu) \). Thus, the only unknown is \( l(0) \).

In the following Proposition, we characterize the equilibrium by letting \( \varphi \) be a conjecture of the equilibrium \( l(0) \). For any conjecture \( \varphi \) of \( l(0) \), Equation (6) provides a conjecture of \( l(n) \), which determines a rational strategy (i.e. join iff \( l(n) > L(n, \mu) \), see Equation (5)). This rational structure is an equilibrium if the likelihood ratio it induces at the empty queue (obtained from Lemma 1) is equal to \( \varphi \). We refer to this fixed point as \( \hat{\varphi} \). In that case, the strategy profile is rational (Definition 1(i) is satisfied) and the corresponding belief is consistent (Definition 1(ii) is satisfied for all \( n \geq 0 \)).

The following Proposition provides further structure to the agent joining equilibrium:

**PROPOSITION 1.** For a given \( \mu \), let \( \hat{n} = (\hat{n}_0, \hat{\alpha}_0, \hat{n}_1) \) where \( \hat{n}_0 \in \mathbb{N} \), \( \hat{\alpha}_0 \in [0, 1) \) and \( \hat{n}_1 \in \mathbb{N} \) (\( \hat{n}_1 > \hat{n}_0 \)) and

\[
\Psi (\hat{n}, \mu) = \frac{\sum_{k=0}^{\hat{n}_0} \left[ \frac{\Delta}{\mu_l} \right]^k + \hat{\alpha}_0 \sum_{k=\hat{n}_0+1}^{\hat{n}_1} \left[ \frac{\Delta}{\mu_h} \right]^k}{\sum_{k=0}^{\hat{n}_0} \left[ \frac{\Delta}{\mu_l} \right]^k + \hat{\alpha}_0 \sum_{k=\hat{n}_0+1}^{\hat{n}_1} \left[ \frac{\Delta}{\mu_h} \right]^k},
\]

(7)

then,

(i) there exists a ‘classical’ threshold equilibrium (pure or mixed if) for some \( \hat{\varphi} \in \mathbb{R}^+ \), and \( \hat{n}_1 = \hat{n}_0 + 1 \) the following conditions are satisfied:

\[
\hat{\varphi} = \Psi ((\hat{n}_0, \hat{\alpha}_0, \hat{n}_0 + 1), \mu)
\]

(8)

and

\[
\begin{cases}
\hat{\varphi} \left( \frac{\mu_l}{\mu_h} \right)^n \geq L(n, \mu), \text{ for } 0 \leq n < \hat{n}_0, \\
\hat{\varphi} \left( \frac{\mu_l}{\mu_h} \right)^n < L(n, \mu), \text{ for } n \geq \hat{n}_0.
\end{cases}
\]

(9)

where the first inequality is binding when \( \alpha_0 > 0 \).

(ii) there exists a non-threshold equilibrium if for some \( \hat{\varphi} \in \mathbb{R}^+ \), and \( \hat{n}_1 > \hat{n}_0 + 1 \) the following conditions are satisfied: the condition of Equation (8) and

\[
\begin{cases}
\hat{\varphi} \left( \frac{\mu_l}{\mu_h} \right)^n > L(n, \mu), \text{ for } 0 \leq n < \hat{n}_0 \text{ (if } 0 < \hat{n}_0), \\
\hat{\varphi} \left( \frac{\mu_l}{\mu_h} \right)^{\hat{n}_0} = L(\hat{n}_0, \mu), \text{ for } n = \hat{n}_0, \\
\hat{\varphi} \left( \frac{\mu_l}{\mu_h} \right)^n > L(n, \mu), \text{ for } \hat{n}_0 < n < \hat{n}_1, \\
\hat{\varphi} \left( \frac{\mu_l}{\mu_h} \right)^n < L(n, \mu), \text{ for } n \geq \hat{n}_1.
\end{cases}
\]

(10)
Proposition 1(i) provides conditions for a pure or mixed strategy agent equilibria: When $\hat{\alpha}_0 = 0$, a pure strategy equilibrium exists in which agents join at any queue length that is strictly less than the threshold $\hat{n}_0$, and balk from queues equal to or longer than $\hat{n}_0$. When $\hat{\alpha}_0 > 0$, and $\hat{n}_1 = \hat{n}_0 + 1$, agents join at any queue length that is strictly less than the threshold $\hat{n}_0$, join with probability $\hat{\alpha}_0$ a queue with length $\hat{n}_0$, otherwise, they balk. This is an extension of a pure strategy equilibrium, as discussed in Hassin and Haviv (2003), p. 7-8. We refer to this equilibrium as the classical threshold equilibrium. Proposition 1(ii) indicates that this classical equilibrium strategy is not always an equilibrium. When $\hat{n}_1 > \hat{n}_0 + 1$, Proposition 1(ii) identifies a non-threshold equilibrium with mixing (with probability $\hat{\alpha}_0$) at some queue length ($\hat{n}_0$), while joining at longer queues and balking at queue length $\hat{n}_1$.

In this following three subsections, we elaborate further the intuition behind Proposition 1. To that end, we introduce $\nu(\varphi, \mu)$, which is the set of solutions to the $\varphi(\mu/\mu_h)^\nu = L(\nu, \mu)$. This is the continuous version of Equation (9), determining $\hat{n}_0$. Define now $\hat{n}(\varphi, \mu) = \lfloor \min \nu(\varphi, \mu) \rfloor$. $\hat{n}(\varphi, \mu)$ is the lowest queue length at which the agent has negative expected utility, and hence balks when the conjecture of the likelihood ratio at the empty queue is $\varphi$ and the service rates are $\mu$.

Keeping all else equal, a larger $\varphi$ means a higher likelihood ratio at the empty queue, and, with Equation (9), a higher likelihood ratio at every queue length. Therefore, the balking threshold $\hat{n}(\varphi, \mu)$ is non-decreasing in $\varphi$. Now, define:

$$\Phi(\hat{n}, \mu) \doteq \Psi\left((\hat{n}, 0, \bullet), \mu\right),$$

(11)

i.e. the likelihood ratio at the empty queue when the customer joining strategy is a pure threshold strategy determined by balking at $\hat{n}$. As $\hat{n}(\varphi, \mu) \in \mathbb{N}$, the function $\Phi(\hat{n}(\varphi, \mu), \mu)$ is a (discontinuous) staircase function in $\varphi$. In the following subsections, $\Phi(\hat{n}(\varphi, \mu), \mu)$ will be discussed further and provide intuition for the equilibrium structure of Proposition 1.

3.1. A pure classical threshold agent joining equilibrium at $\hat{n}_0$ with $\hat{\alpha}_0 = 0$

A pure strategy threshold joining equilibrium (with $\hat{\alpha}_0 = 0$) exists if there is a solution, $\hat{\varphi}$, to:
\[ \varphi = \Phi(\hat{n}(\varphi, \mu), \mu). \] (12)

In words, if the conjecture about the likelihood ratio at the empty queue, \( \hat{\varphi} \), leads to a rational joining strategy with threshold \( \hat{n}_0 = \hat{n}(\hat{\varphi}, \mu) \), which in turn leads to a likelihood ratio at the empty queue, \( \Phi(\hat{n}_0, \mu) \) that is consistent with (i.e. equal to) \( \hat{\varphi} \), then the two equilibrium conditions of Definition 1(i) and (ii) are satisfied. \( \hat{n}_1 \) is irrelevant as it is reached with zero probability.

**Example of a pure threshold agent joining equilibrium:** Now, we provide an illustrative example of a case when no threshold agent joining equilibrium exists. Let \( v_h = 3, v_l = -1, p = 0.5, c = 0.2505, \Lambda = 1 \) and \( \mu = 1.05 \) and \( \overline{\mu} = 1.1 \). In Figure 1, left panel, we illustrate \( \Phi(\hat{n}(\varphi, \mu_{sf}), \mu_{sf}) \), i.e. the likelihood ratio when the high quality firm’s service rate is slower than the low quality firm’s service rate. We also indicate the value of \( \hat{n}(\varphi, \mu_{sf}) \) on each horizontal segment where \( \hat{n}(\varphi, \mu_{sf}) \) is constant. Note that at \( \varphi = 0 \), the agent’s balking threshold is zero as the low quality service value is negative (\( v_l = -1 \)). It follows that the queue is always empty, irrespective of the firm’s quality.

Hence: \( \Phi(0, \mu) = 1 \) for \( \varphi = 0 \) and for low values of \( \varphi \). As \( \varphi \) increases, \( \hat{n}(\varphi, \mu_{sf}) \) increases, as indicated in the Figure. For the range of values of \( \varphi \) for which \( \hat{n} = 4 \) includes \( \Phi(4, \mu_{sf}) \). Hence, we obtain that \( \hat{\varphi}(\mu_{sf}) = 0.9172 \) is a fixed point of Equation (12) and \( \hat{n}_{sf} = 4 \). Notice that \( \Phi(\hat{n}(\varphi, \mu_{sf}), \mu_{sf}) \) decreases. As a result, there can be at most one pure strategy equilibrium. On the right panel, the equilibrium expected service value and waiting cost are plotted for a randomly arriving agent, when assuming that all other agents balk at a queue length of 4. Indeed, it is for a randomly arriving agent also rational to balk at a queue length of 4.

Furthermore, notice that both the expected value and the expected costs increase as a function of the queue length. This is because the high quality firm is slow under \( \mu_{sf} \). Finally, notice that \( \hat{\varphi}(\mu_{sf}) < 1 \). Recall that the equilibrium belief, \( \hat{\gamma}(0) \) satisfies \( \hat{\gamma}(0)/(1 - \hat{\gamma}(0)) = l(0) \times l_0 \), where \( l_0 = p/(1 - p) \) reflects the prior belief and, in equilibrium, \( l(0) = \hat{\varphi}(\mu_{sf}) \) is the likelihood ratio of observing the empty queue. Hence, when \( \hat{\varphi}(\mu_{sf}) < 1 \), upon observing the empty queue, the posterior belief that the firm is of high quality is lower than the prior belief. In other words, when
the high quality firm’s service rate is lower than the low quality firm’s service rate, an empty queue is ‘bad’ news. This is intuitive as rational agents associate high quality with long queues.

Assume now that at one of the discontinuous points of \( \hat{n}(\varphi, \mu) \), say, \( \hat{\varphi} \), the agents’ threshold changes from \( \hat{n}_0 \) to \( \hat{n}_1 \) and causes the likelihood ratio induced by the threshold to change from a level strictly below (or: above) \( \hat{\varphi} \) to a level strictly above (or: below) \( \hat{\varphi} \). Formally, when there exists a \( \hat{\varphi} \) such that \( \hat{n}(\hat{\varphi} - \epsilon, \mu) = \hat{n}_0 \) and \( \hat{n}(\hat{\varphi} + \epsilon, \mu) = \hat{n}_1 \) for which \( \Phi(\hat{n}_0, \mu) < \Phi(\hat{n}_1, \mu) \) (or: \( \Phi(\hat{n}_0, \mu) > \Phi(\hat{n}_1, \mu) \)). When \( \hat{\varphi} \in (\Phi(\hat{n}_0, \mu), \Phi(\hat{n}_1, \mu)) \) (or: \( \hat{\varphi} \in (\Phi(\hat{n}_1, \mu), \Phi(\hat{n}_0, \mu)) \)), \( \hat{\varphi} \) cannot characterize a pure strategy equilibrium. In that case, we can construct an agent joining equilibrium involving mixing at \( \hat{n}_0 \), i.e. \( \alpha_0 > 0 \).

We consider two cases involving mixing in the following two subsections. In the first case, we analyze \( \hat{n}_1 = \hat{n}_0 + 1 \). This corresponds with a classical threshold joining equilibrium with randomization. In the second case, \( \hat{n}_1 > \hat{n}_0 + 1 \). We will denote the latter as ‘sputtering’ equilibrium, because at \( \hat{n}_0 \) the queue ‘sputters’ before increasing to \( \hat{n}_1 \) due to the mixing probability \( \hat{\alpha}_0 \) at \( \hat{n}_0 \).
3.2. Classical threshold agent joining equilibrium at $\hat{n}_0$ with randomization $\hat{\alpha}_0 > 0$ and $\hat{n}_1 = \hat{n}_0 + 1$

When $\hat{n}_1 = \hat{n}_0 + 1$, it is easy to see from the definitions of Equations (7) and (11) that:

$\Psi((\hat{n}_0,0,\hat{n}_0+1), \mu) = \Phi(\hat{n}_0, \mu)$ and $\Psi((\hat{n}_0,1,\hat{n}_0+1), \mu) = \Phi(\hat{n}_0+1, \mu)$. Then, by continuity of $\Psi((\hat{n}_0, \alpha, \hat{n}_0+1), \mu)$ in $\alpha$, any likelihood ratio in $[\Phi(\hat{n}_0, \mu), \Phi(\hat{n}_0+1, \mu)]$ is achieved with a mixed strategy, $\alpha \in [0,1]$ at $\hat{n}_0$. Thus, there exists an equilibrium with mixing at $\hat{n}_0$ such that the likelihood ratio of that strategy is equal to $\hat{\varphi}$. Both conditions (i) and (ii) of Definition 1 are satisfied.

It follows that a classical mixed strategy extension of a pure strategy threshold (Hassin and Haviv, 2003) is an equilibrium; i.e. all agents join queues that are strictly shorter than $\hat{n}_0$, randomize with probability $\hat{\alpha}_0$ at a queue of length $\hat{n}_0$, and balk at queues that are longer than or equal to $\hat{n}_0 + 1$.

Example of threshold agent joining equilibrium with randomization: In Figure 2, left panel, we illustrate $\Phi(\hat{n}(\varphi, \mu_{fs}), \mu_{fs})$ for the same parameter values as for Figure 1, but, for a different (given) service rate strategy, $\mu_{fs}$, i.e. when the high quality firm is faster than the low quality firm. Again, we plot the values of $\hat{n}(\varphi, \mu_{fs})$. Note that $\Phi(\hat{n}(\varphi, \mu_{fs}), \mu_{fs})$ increases. Observe that two fixed points of Equation (12) are possible: one at $\hat{n} = 3$ and one at $\hat{n} = 4$. In addition,
at $\hat{\varphi} = 1.0755$: $\hat{n}_0 = 3$ and $\hat{n}_1 = 4$ and observe from the Figure that $\hat{\varphi} \in (\Phi(\hat{n}_0, \mu), \Phi(\hat{n}_1, \mu))$. Hence, there is also one mixed strategy equilibrium possible with $\hat{\varphi}(\mu_{fs}) = 1.0755$, $\hat{n}_{fs} = 3$ and $\hat{\alpha}_0 = 0.3081$. On the right panel, the equilibrium expected service value and waiting cost are plotted for a randomly arriving agent, when assuming that all agents play the randomized strategy (at 3). Notice that the randomly arriving agent is also indifferent between joining and not joining at queue length 3, and at strictly lower queue lengths, he prefers joining. Furthermore, notice that the expected value decreases and the expected cost increases as a function of the queue length. This is because the high quality firm is fast under $\mu_{sf}$. A longer queue is bad news in two ways: it implies more waiting and also a service of less value. Finally, notice that it is intuitive that $\hat{\varphi}(\mu_{sf}) > 1$, i.e. an empty queue is ‘good’ news (the prior increases upon observing an empty queue because high quality is associated with fast service).

3.3. Non-threshold or ‘sputtering’ agent joining equilibrium with $\hat{\alpha}_0 > 0$ and $\hat{n}_1 > \hat{n}_0 + 1$

When $\hat{n}_1 > \hat{n}_0 + 1$, an extension of a pure threshold strategy with mixing with probability $\alpha_0$ at the threshold $\hat{n}_0$ and balking at $\hat{n}_0 + 1$ cannot cover any likelihood ratio in $[\Phi(\hat{n}_0, \mu), \Phi(\hat{n}_1, \mu)]$ when $\Phi(\hat{n}_0 + 1, \mu) < \Phi(\hat{n}_1, \mu)$ and $\hat{\varphi} \in (\Phi(\hat{n}_0 + 1, \mu), \Phi(\hat{n}_1, \mu)]$. In that case, mixing at $\hat{n}_0$ only covers the likelihood ratio range of $[\Phi(\hat{n}_0, \mu), \Phi(\hat{n}_0 + 1, \mu)]$ and leaves the range $(\Phi(\hat{n}_0 + 1, \mu), \Phi(\hat{n}_1, \mu)]$ uncovered. If $\hat{\varphi}$ lies in the latter range, the classical mixed strategy extension cannot determine the equilibrium.

Proposition 1(ii) identifies a non-threshold strategy. By continuity of $\Psi((\hat{n}_0, \alpha, \hat{n}_1), \mu)$ in $\alpha$, any likelihood ratio in $[\Phi(\hat{n}_0, \mu), \Phi(\hat{n}_1, \mu)]$ is achieved with a mixed strategy, $\alpha \in [0, 1]$ at $\hat{n}_0$ (as can be seen from the definitions of Equations (7) and (11)). Hence, there exists a randomization probability $\hat{\alpha}_0$ that $\hat{\varphi}$ is reached if it lies in $(\Phi(\hat{n}_0 + 1, \mu), \Phi(\hat{n}_1, \mu)]$.

Example of a non-threshold or sputtering agent joining equilibrium: Now, we provide an illustrative example of a case when no threshold agent joining equilibrium exists. Let $v_h = 2$, $v_l = 0$, $p = 0.1$, $c = 0.051$ and $\mu = 1.05$ and $\overline{\mu} = 1.25$. In Figure 3, left panel, we illustrate $\Phi(\hat{n}(\varphi, \mu_{sf}), \mu_{sf})$. We indicate again the value of $\hat{n}(\varphi, \mu_{sf})$ on each horizontal segment where $\hat{n}(\varphi, \mu_{sf})$ is constant.
Notice that at $\hat{\varphi} = 0.5439$, $\hat{n}$ increases from $\hat{n}_0 = 6$ to $\hat{n}_1 = 40$. We obtain a non-threshold equilibrium with $\hat{\varphi}(\mu_{sf}) = 0.5439$, $\hat{n}_0 = 6$ and $\hat{n}_1 = 40$. This means that agents always join the queue as long as its length is less than 6, they join with probability 0.1170 when the queue length is 6, they always join when the queue length is between 7 and 39, and they balk from any queue that is 40 or longer.

Now, we explain the intuition behind this result: Recall that $\nu(\varphi, \mu)$ is the set of real roots of $\varphi(\mu_t/\mu_h) = L(\nu, \mu)$. $[\nu(\varphi, \mu_{sf})]$ determines the queue length at which rational agents do not join, assuming that $\varphi$ is the likelihood ratio at the empty queue. On the left panel, we plot in dashed lines the correspondence $\Phi([\nu(\varphi, \mu_{sf})], \mu_{sf})$, where for certain values of $\varphi$, $\nu(\varphi, \mu_{sf})$ can take three different values. The values of $[\nu(\varphi, \mu_{sf})]$ over the dashed branch are between 7 and 40.

Recall also that $\hat{n}(\varphi, \mu_{sf}) = [\min \nu(\varphi, \mu_{sf})]$, i.e. the lowest queue length at which an agent balks. Hence, whenever there are multiple solutions in $\nu(\varphi, \mu_{sf})$, the solution on the solid, top branch is selected. As a result, the range of queue lengths between 7 and 39 is excluded from $\hat{n}(\varphi, \mu)$, which ‘jumps’ from 6 to 40, see Figure 3, left panel. As $\hat{\varphi} \in (\Phi(7, \mu_{sf}), \Phi(40, \mu_{sf}))$, no classical threshold equilibrium can exist. This can also be observed from the Figure as the 45 degree line has no intersection with $\Phi(\hat{n}(\varphi, \mu_{sf}) \mu_{sf})$. Thus, with Proposition 1(ii), a non-threshold equilibrium is identified with randomization at a queue of length 6 and balking at a queue of length 40. The right panel of Figure 3 illustrates the equilibrium agent updated utility (value and waiting cost). Indeed, there is one queue length (6), at which the agent is indifferent between joining and balking, while at queue length 40, the agent balks.

It is interesting to observe that a non-threshold equilibrium causes the following queue dynamics:

As at $\hat{n}_0$, agents join the queue with a probability of less than one, thus it will be difficult for the queue to grow beyond $\hat{n}_0$. But, once the queue is larger than $\hat{n}_0$, all agents join again with probability one. Hence, the firm will observe that the queue ‘stalls’ from time to time at $\hat{n}_0$. This is why we label this equilibrium as a ‘sputtering’ equilibrium. These queue dynamics are an immediate result of the non-threshold agent strategy. In the following subsection, we discuss when this phenomenon occurs.
Figure 3 Determination of equilibrium; $\Phi (\hat{n}(\phi, \mu_{sf}), \mu_{sf}) = \phi$ (right panel) and the Updated Service Value, $\hat{\gamma}(n)v_h + (1 - \hat{\gamma}(n))v_l$, vs. the Updated Waiting Cost, $\hat{\gamma}(n)c(n+1)/\mu_h + (1 - \hat{\gamma}(n))c(n+1)/\mu_l$ (left panel).

Demonstration of a non-threshold equilibrium with randomization at $n = 6$ and a balking threshold at $n = 40$.

3.4. Properties of the agent joining equilibrium

In this subsection, we discuss several properties of the agent joining equilibrium. First, we derive a sufficient condition for the classical threshold agent joining equilibrium (Lemma 2). Next, we discuss the number of possible agent joining equilibria (Lemma 3). Finally, we discuss in Lemma 4 and in Proposition 2 some important comparative statics of the threshold of the equilibrium.

From the discussion in the previous subsection, it follows that it is important to assess when $\hat{n}(\phi, \mu)$ increases in steps of +1, or if it has larger jumps. When $(\mu_h/\mu_l)^\nu L(\nu, \mu)$ is monotone in $\nu$ and $\nu(\phi, \mu)$ is a singleton, then, obviously, $\hat{n}(\phi, \mu)$ increases in steps of +1. In general, however, $(\mu_h/\mu_l)^\nu L(\nu, \mu)$ is not always monotone in $\nu$.

$\nu(\phi, \mu)$ can contain three solutions, as illustrated in Figure 3, left panel, where for a range of values of $\phi$ two solutions are indicated by means of a dashed line and one solution (the lowest value) by means of the solid line. The reason why there are three possible solutions could be understood from Figure 3: As the service value is bounded between $v_l$ and $v_h$, the updated service value after observing the queue length follows an S-shape for this particular example: the updated value is bounded by $v_h$ and $v_l$ and at very low and very high queue lengths, finding one extra agent in the queue provides low additional evidence that the firm is of high quality. The updated waiting
cost, on the other hand, is monotone increasing. As a result, there may be three intersection points between the S-shaped updated value and the monotone increasing updated cost. Hence, \( \hat{n}(\phi, \mu) \) can increase in steps that are larger than +1 (in Figure 3, from 6 to 40 at \( \hat{\phi} = 0.5439 \)). The following Lemma provides a sufficient condition for \( \hat{n}(\phi, \mu) \) to be increasing in steps of +1:

**Lemma 2.** When \( C(\mu) > 0 \), where

\[
C(\mu) = c + \frac{1}{4} (\mu_h v_h - \mu_\ell v_\ell) \ln \left( \frac{\mu_h}{\mu_\ell} \right),
\]

then, \( \hat{n}(\phi, \mu) \) increases in steps of +1, otherwise, \( \hat{n}(\phi, \mu) \) may increase in steps that are larger than +1.

It follows from Lemma 2 that \( C(\mu_{fs}) = c + \frac{1}{4} (\mu v_h - \mu v_\ell) \ln \left( \frac{\mu v_h}{\mu v_\ell} \right) \) is strictly positive (as \( \mu v_h > \mu v_\ell \) and \( \mu / \mu > 1 \)) and, trivially, \( C(\mu_{ss}) = C(\mu_{ff}) = c \) is strictly positive. Hence, when the firm’s service rate is independent of the firm’s quality, or, when the highest quality firm is also the fastest firm, threshold joining strategies always exist as the updated service value is either independent of the queue length, or decreases in queue length, while the updated service cost increases in the queue length. Only when the highest quality firm is also the slowest and both service value and cost increase in the queue length, a sputtering equilibrium may exist. As a consequence, if \( C(\mu_{sf}) = c + \frac{1}{4} (\mu v_h - \mu v_\ell) \ln \left( \frac{\mu v_h}{\mu v_\ell} \right) > 0 \), a classical threshold strategy is an agent joining equilibrium for any pure service rate strategy, \( \mu_\sigma \) with \( \sigma \in \{ ss, ff, sf, fs \} \). In terms of the primitives of our model: If the difference between the slow (\( \mu \)) and the fast speed (\( \mu \)) is not high, or, that the waiting costs, \( c \), are high enough, provided that the high quality firm operating at a slow speed creates more value per unit of service time than the low quality firm at a fast service speed (\( \mu v_h > \mu v_\ell \)), the agent joining equilibrium is a classical threshold strategy for any service rate strategy.

**Illustration:** The condition \( C(\mu) = 0 \) is illustrated in Figure 4. The dashed lines are \( \mu_h v_h = \mu_\ell v_\ell \) and \( \mu_h = \mu_\ell \). Notice that \( C(\mu) < 0 \) only when \( \mu_h v_h > \mu_\ell v_\ell \) and \( \mu_h < \mu_\ell \), i.e., for a given service rate of the low quality firm; when the service rate of the high quality firm is lower, but, not too low, it is possible that \( C(\mu) < 0 \). \( C(\mu) > 0 \) is a sufficient (but not necessary) condition for a threshold equilibrium strategy.
Non-threshold Equilibrium may exist
Threshold Equilibrium always exists
\[ C(\mu) < 0 \]
\[ C(\mu) > 0 \]

Figure 4 \( C(\mu) = 0 \) for \( \nu_h = 1.75, \nu_L = 0.25 \) and \( c = 0.1051 \). The dashed lines are \( \mu_h\nu_h = \mu_L\nu_L \) and \( \mu_h = \mu_L \).

Threshold Equilibria: When \( C(\mu) > 0 \), we analyze the properties of threshold strategy equilibria, i.e. when \( \hat{\alpha}_0 = 0 \) or \( \hat{\alpha}_0 > 0 \) and \( \hat{n}_1 = \hat{n}_0 + 1 \). In the next Lemma, we characterize the number of pure strategy joining equilibria:

**Lemma 3.** When \( C(\mu) > 0 \), then:

(i) When \( \mu_h < \mu_L \), there exists at most one pure strategy threshold equilibrium. If no pure strategy equilibrium exists, a mixed strategy equilibrium exists.

(ii) When \( \mu_L \leq \mu_h \), there exists at least one pure strategy threshold equilibrium. If more than one pure strategy equilibrium exists, a mixed strategy equilibrium exists.

The intuition behind Lemma 3 is the following: Consider the situation when the high quality firm is the slowest firm \((\mu_h < \mu_L)\); it is thus easily understood that the likelihood ratio at the empty queue, \( \Phi(\hat{n}, \mu) \), decreases in the joining threshold, \( \hat{n} \): Assume that \( \hat{n} \) is very low. The queue is mostly short. Hence, the recurrent (queue length) state space is small such that the probability of observing an empty queue when the product quality is high is not different from the probability of observing an empty queue when the product quality is low. Hence, \( \Phi \) is about 1. When \( \hat{n} \) is high, there is enough variation in the queue lengths such that there is a difference in probability of observing an empty queue. In fact, the probability of observing an empty queue when the quality
is high will be lower than the probability of observing an empty queue when the quality is low. Hence, $\Phi$ is less than 1 and $\Phi$ is thus decreasing in $\hat{n}$. As $\hat{n}(\varphi, \mu)$ is always increasing in $\varphi$, it follows that $\Phi(\hat{n}(\varphi, \mu), \mu)$ decreases in $\varphi$. This was the case in Figure 1. As a result, there can be at most one point of intersection for $\Phi(\hat{n}(\varphi, \mu), \mu)$ and the 45 degree line, and, at most, there is one pure strategy equilibrium. When the low quality firm is slow, the opposite is true: $\Phi(\hat{n}, \mu)$ increases. As a result, there can be multiple points of intersection for $\Phi(\hat{n}(\varphi, \mu), \mu)$ and the 45 degree line, and thus there can be multiple pure strategy equilibria. This was the case in Figure 2.

In summary, combining Proposition 1 and Lemmas 2 and 3 yield the following insights:

- When the high quality firm is fast and the low quality firm is slow, long queues are ‘bad news’. Arriving at an empty queue ‘boosts’ the confidence in the customer’s mind about the value of service. The updated value decreases as the queue gets longer. On the other hand, the wait cost also increases as the queue gets longer. Therefore, a threshold joining policy is an equilibrium.

- When the high quality firm is slow and the low quality firm is fast, long queues are ‘good news’. The updated value increases in the queue length. Long queues also imply long waiting. Arriving at an empty queue reduces the confidence that the quality is high. Hence, both costs and value increase. When the difference between the service rates of the high- and the low-quality firm is small, there will still exist a threshold, at which the additional updated value does not compensate for the increase of the waiting costs.

- When the difference in service rates is large enough, it may be problematic to join short queues. In this case, a “sputtering” equilibrium involving randomization at some queue length emerges. Agents are indifferent between joining or not at that queue length, while, at strictly longer and possibly shorter queue lengths, the agents may strictly prefer joining. As a result, the queue generally resides at low lengths, and probabilistically (depending on the random decision of one agent) grows to a much larger queue, since every one else joins at higher queue lengths.

Now, we can state the following property of the thresholds:

**Lemma 4.** If a pure strategy equilibrium, $\hat{n}_{sf}$ and $\hat{n}_{fs}$ exists if $v_h = v_l = v$, then $\hat{n}_{ss} = \left\lfloor v\mu/c \right\rfloor$, $\hat{n}_{ff} = \left\lfloor v\mu/c \right\rfloor$, $\hat{n}_{sf} = \left\lfloor v(p/\mu + (1 - p)/\mu)^{-1}/c \right\rfloor$ and $\hat{n}_{fs} = \left\lfloor v(p/\mu + (1 - p)/\mu)^{-1}/c \right\rfloor$. 

**Debo and Veeraraghavan: Unknown Service Quality and Speed**
When there is only uncertainty about the service rate (i.e. when \( v_h = v_\ell \)), the agents ‘average’ the waiting costs associated with the different service rates ((\( p/\mu + (1 - p)/\bar{\mu} \))\(^{-1} \) for \( \mu_{sf} \) or (\( p/\bar{\mu} + (1 - p)/\mu \))\(^{-1} \) for \( \mu_{fs} \)). As a result, the average waiting cost is between the extreme realizations (\( \mu \) and \( \bar{\mu} \)). The joining threshold with a service strategy, \( \mu_\sigma, \sigma \in \{sf, fs\} \) is between the joining threshold for service strategies, \( \mu_\sigma, \sigma \in \{ss, ff\} \). In other words, unobservable service rates lead to queue joining behavior that is the ‘average’ of the queue joining behavior assuming that the firm is either fast or slow.

**Proposition 2.** If a pure strategy equilibrium, \( \hat{n}_{sf} \) and \( \hat{n}_{fs} \), exists:

There exist a \( \mu_{sf}^* \) and \( \mu_{fs}^* \) such that for \( pv_h + (1 - p)v_\ell = v \), then \( \hat{n}_{ss} = [v\mu/c], \hat{n}_{ff} = [v\bar{\mu}/c] \) and

(i) \( \hat{n}_{sf} \geq \hat{n}_{ff} \) when \( \mu \in [\mu_{sf}^*, +\infty) \) (and \( \mu \leq \bar{\mu} \)) and

(ii) \( \hat{n}_{ff} \geq \hat{n}_{fs} \) when \( \bar{\mu} \in [\mu_{fs}^*, +\infty) \) (and \( \mu \leq \mu \)).

Proposition 2 reveals the fundamental impact on the agent strategy of quality uncertainty (\( v_h > v_\ell \)). When there is no quality uncertainty (\( v_h = v_\ell = v \)) notice that \( \hat{n}_{sf} \) is increasing in \( \mu \). But, when there is quality uncertainty (\( v_h > v_\ell \), keeping \( pv_h + (1 - p)v_\ell = v \)) the joining thresholds \( \hat{n}_{sf} \) decreases in \( \mu \) when the high quality firm’s service rate is high enough. It can even be the case that \( \hat{n}_{sf} \) is greater than \( \hat{n}_{ff} \) when \( \mu \) is low enough. In Table 1, we provide a numerical illustration of this observation (on the third and fourth row). Note that this would never occur when there is no uncertainty about the quality. This is because when the high quality firm is slow, the inconvenience of long waiting times is overcome by the increased confidence that the service quality is high. This comparative static is only valid when the high quality firm’s service rate is high enough, because otherwise congestion effects dominate the value effect.

Similarly, it is worth noting that when there is no quality uncertainty (\( v_h = v_\ell = v \)), \( \hat{n}_{fs} \) increases in \( \bar{\mu} \), but when there is quality uncertainty (\( v_h > v_\ell \) and \( pv_h + (1 - p)v_\ell = v \)) the joining threshold \( \hat{n}_{fs} \) decreases in \( \bar{\mu} \). We also illustrate in Table 1 that it can even be the case that \( \hat{n}_{fs} \) is lower than \( \hat{n}_{ss} \) when \( \bar{\mu} \) is high enough (on the third row). This is because when the high quality firm is fast, there are two reasons for the agent not to join a long queue: high waiting cost and low updated
Table 1 Illustration of classical threshold equilibria for different service rate strategies ($\sigma \in \{ss, fs, sf, ff\}$). All other parameters are: $p = 1/2$, $c = 0.151$, $r = 1$ and $\Lambda = 1$. Note that for all cases $pv_h + (1-p)v_l = 1$.

<table>
<thead>
<tr>
<th>$v_l$</th>
<th>$v_h$</th>
<th>$\mu$</th>
<th>$\tilde{\mu}$</th>
<th>$n_{ss}$</th>
<th>$n_{fs}$</th>
<th>$n_{sf}$</th>
<th>$n_{ff}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1.050</td>
<td>1.250</td>
<td>6</td>
<td>7</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0.750</td>
<td>0.925</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>-0.85</td>
<td>2.85</td>
<td>1.050</td>
<td>1.250</td>
<td>6</td>
<td>4</td>
<td>16</td>
<td>8</td>
</tr>
<tr>
<td>-0.85</td>
<td>2.85</td>
<td>0.750</td>
<td>0.925</td>
<td>4</td>
<td>4</td>
<td>7</td>
<td>6</td>
</tr>
</tbody>
</table>

4. When to Slow Down Service?

So far, we kept the service rate $\mu$ or the firm’s strategy, $\beta$, fixed. Now, we study the equilibrium service rate selection by imposing the rationality condition of Definition 1(iii). To keep focus, we assume in this section that $C(\mu_{sf}) > 0$. From Lemma 2, all agent equilibria (satisfying conditions (i) and (ii) of Definition 1) are classical threshold equilibria. Define now (with a slight abuse of notation):

$$R(n, \alpha, \mu) = r\{\sum_{k=0}^{n-1} \Lambda \pi(k, \alpha, \mu) + \alpha \Lambda \pi(n, \alpha, \mu)\}$$

But still assuming pure threshold equilibria, see Lemma 2.
where \( \alpha(k) = 1 \) for \( 0 \leq k < n \), \( \alpha(n) = \alpha \) and \( \alpha(k) = 0 \) for \( n + 1 \leq k \). \( R(n, \alpha, \mu) \) is the revenue rate of the firm when the agent strategy is characterized by a threshold at \( n \) and mixing with probability \( \alpha \) and define

\[
\Delta(n, \alpha) = R(n, \alpha, \mu) - R(n, \alpha, \mu).
\]

\( \Delta(n, \alpha) \) is the increase in revenue rate when the firm changes from the slow to the fast service speed, when the threshold is determined by \( n \) and when there is possible mixing of \( \alpha \) at \( n \). Finally, define \( \Delta(\sigma) = \Delta(\hat{n}_\sigma, \hat{\alpha}_\sigma) \), which is the increase in revenue rate when speeding up under strategy \( \sigma \in \{ss, ff, sf, fs\} \). With Definition 1(iii), for a given rational agent joining strategy, \( \alpha^* \), and a consistent belief, \( \gamma^* \), \( \beta^* \) must be such that no firm has an incentive to deviate from \( \beta^* \). The increase in revenue rate of speeding up is given by \( \Delta(\sigma) \), and the cost of speeding up for the \( \omega \) type firm is given by \( k_\omega \).

To make the notation easier, we denote the cost difference between the high and low quality firm, \( k_h - k_\ell \), by \( \delta_k \) and write \( k \) for \( k_\ell \). We say that \( k \) is the base cost of speeding up. The following Lemma characterizes \( \beta^* \):

**Lemma 5.** (i) \( k < \Delta_{ff} - \delta_k : \beta^* = (1, 1) \), \( \Delta_{sf} - \delta_k < k < \Delta_{ss} : \beta^* = (0, 1) \) and \( \Delta_{ss} < k : \beta^* = (0, 0) \),

(ii) \( \min(\Delta_{sf}, \Delta_{ss}) < k < \max(\Delta_{ss}, \Delta_{sf}) : \beta^* = (0, \beta^*_h(k)) \) and \( \min(\Delta_{sf}, \Delta_{ff}) - \delta_k < k < \max(\Delta_{ff}, \Delta_{sf}) - \delta_k : \beta^* = (\beta^*_h(k), 1) \).

Lemma 5 follows immediately from Equation (2) in Definition 1(iii). Lemma 5 verifies for each pure strategy service rate equilibrium whether the optimality conditions are satisfied. For the mixed strategy service rate equilibrium, Lemma 5 determines the indifference conditions of the mixing firm. Note that \( \mu_{fs} \) is never an equilibrium. This is because the cost of capacity for the high quality firm is higher than or equal to the cost of capacity for the low quality firm, while the revenues for both firms are the same.

When the cost of speeding up for the most expensive firm (the high quality firm) is less than the increase in revenue rate, both firms speeding up is an equilibrium. When the cost of speeding up of the least expensive firm (the low quality firm) is less than the increase in revenue rate, then,
both firms slowing down is an equilibrium. Otherwise, the low quality firm speeds up and the high quality firm does not speed up. That is captured in Lemma 5(i). This is intuitive. In addition, Lemma 5(ii) also identifies mixed strategy equilibria in which one firm randomizes its service rate. Notice from Lemma 5(ii) that in one equilibrium, \( \beta_h^*(k) \leq \beta_l^* = 1 \) and \( \beta_h^* = 0 \leq \beta_l^*(k) \) in the other equilibrium, i.e. in all pure or mixed strategy equilibria, the service rate of the high quality firm is never larger than the service rate of the low quality firm.

In order to find the equilibrium service rate selection, we need to find an ordering for \( \Delta_\sigma, \sigma \in \{ss, ff, sf, fs\} \). Proposition 3 identifies conditions that allow us to rank \( \{\Delta_\sigma\} \):

**Proposition 3.** There exists a unique \( \rho_n < 1 \) that increases in \( n \) and for which \( \lim_{n \to \infty} \rho_n = 1 \), such that for small \( \epsilon > 0 \) for which \( \epsilon > \frac{\mu}{\bar{\mu}} - 1 > 0 \):

(Case-i): When \( \frac{\mu}{\bar{\mu}} < \frac{\mu}{\bar{\mu}} < \rho_n \), then \( \Delta(n, 0) \approx \frac{\partial R(n, 0, \mu)}{\partial \mu} (\bar{\mu} - \mu) \) decreases in \( n \),

(Case-ii): When \( \rho_n < \frac{\mu}{\bar{\mu}} < \frac{\mu}{\bar{\mu}} \), then \( \Delta(n, 0) \approx \frac{\partial R(n, 0, \mu)}{\partial \mu} (\bar{\mu} - \mu) \) increases in \( n \).

According to Proposition 3, the increase in revenues from speeding up (from \( \mu \) to \( \bar{\mu} \)) increases in the (pure strategy) joining threshold \( n \) when the service rates are relatively low, and decrease when the service rates are relatively high. Proposition 3 can be understood intuitively as follows: Assume that the joining threshold, \( n \), is infinitely large, i.e. \( n = \infty \), then, the revenue rate, \( R \), is equal to \( \min(\mu, \Lambda) \): When the service rate is less than the arrival rate, the server is busy 100% of the time and hence, the revenue rate is \( \mu \). When the service rate is more than the arrival rate, all agents join. Hence, the revenue rate is \( \Lambda \). For \( \mu < \Lambda \), the marginal increase of the revenue rate is the highest possible: \( \frac{\partial R}{\partial \mu} = 1 \). Hence, lower values of \( \hat{n} \) (that are finite) must have a lower marginal increase. For \( \mu > \Lambda \), the marginal increase of the revenue rate is the lowest possible: \( \frac{\partial R}{\partial \mu} = 0 \). Hence, lower values of \( \hat{n} \) (that are finite) must have a higher marginal increase. It is thus easily understood that the marginal increase in revenues increases in the joining threshold when the service rate is less than the agent arrival rate, and decreases when the service rate is more than the agent arrival rate.

Figure 5 further illustrates the agent-equilibrium revenues for a firm with service rate \( \mu \) under
two different threshold equilibria with thresholds $\hat{n} = 4$ and $\hat{n} = 7$ and an agent arrival rate $\Lambda$ that is equal to one. As long as the service rate is less than the demand, note that as the joining threshold increases, the marginal return from speeding up (from $\mu = 0.5$ to $\mu = 0.6$) increases in $\hat{n}$ (i.e. $\Delta(\hat{n}, 0)$ increases from 0.0824 for $\hat{n} = 4$ to 0.0951 for $\hat{n} = 7$). When the service rate is more than the market demand $\Lambda$, the marginal return to speeding up (from $\mu = 1.25$ to $\mu = 1.35$) decreases in $\hat{n}$ (i.e. $\Delta(\hat{n}, 0)$ decreases from 0.0213 for $\hat{n} = 4$ to 0.0155 for $\hat{n} = 7$). Proposition 3 has implications for the existence of the equilibria of Proposition 5.

Let $\beta^*$ be the service choice made by the firm. The equilibrium profits of the firm can be written as $\Pi_\omega^* = \beta^*(\omega)(R(\alpha^*, \mu) - k_\omega) + (1 - \beta^*(\omega))R(\alpha^*, \mu)$. We now apply Proposition 3 and Lemma 5 to various parameter settings. The results yield insights into the impact of the quality uncertainty on the firm’s service rate selection and equilibrium profits.

4.1. Uncertainty about the product quality only ($v_h > v_l$ and $k_h = k_l$).

We examine a market with uncertainty about the service value, but not about the costs of speeding up. There is no cost-related reason for the rational agents to expect that the high and low quality firms select a different service rate as the cost difference between the fast and slow service is the same for both firms. We investigate when both firms invest in fast service, and when they persist with slow service. When the base cost is high ($k > \Delta_{ss}$), both firms are slow and when the base
cost is low (k < \Delta_{ff}) both firms are fast. This is intuitive.

Recall from section 3.4 that when the market demand is low, the agents’ balking threshold when one firm is slow, may be higher than the balking threshold when both the firms are fast. This result is in contrast to the classical observable queueing results, where balking thresholds increase with service rates. In effect, we have \hat{n}_{sf} > \hat{n}_{ff} > \hat{n}_{ss} vide Proposition 2 (see the numerical example of the third and fourth rows in Table 1). With Proposition 3, we also obtain a ranking of the marginal revenues from speeding up that depends on the relative value of the service rate versus the potential agent arrival rate:

Figure 6 The equilibrium service strategy (\beta^*, panels (A) and (C)) and profit rates (\Pi^*, panels (B) and (D)) respectively as a function of k for \nu_h > \nu_l and \kappa_h = \kappa_l (for section 4.1). Panels (A) and (B) are for high service rates Panels (C) and (D) are for low service rates, compared to the potential demand arrival rate The dashed lines are the low quality firm’s service strategy and profit rates.
$\Delta_{sf} < \Delta_{ff} < \Delta_{ss}$ when the service rate is high (i.e. when $\Lambda/\mu < \Lambda/\mu < \rho_{ss}$) (case $i$) and $\Delta_{ss} < \Delta_{ff} < \Delta_{sf}$ when the service rate is low (i.e. when $\rho_{sf} < \Lambda/\mu < \Lambda/\mu$) (case $ii$).

As a result, when the service rate is high (case $i$), in the range costs of speeding up, $k \in [\Delta_{sf}, \Delta_{ss}]$, a mixed strategy equilibrium in service rates exists (see Lemma 5). The mixed equilibrium can be classified into two types depending on the base costs $k$:

(i) The high quality firm is slow, and the low quality firm mixes between fast and slow service rates (i.e. $\beta^*(k) = (0, \beta^*_l(k))$).

(ii) The high quality firm mixes between fast and slow service rates, and the low quality firm is fast (i.e. $\beta^*(k) = (\beta^*_h(k), 1)$).

For these mixed strategy equilibria, regardless of the mixing probabilities, note that the high quality firm’s expected service rate is always lower than the low quality firm’s expected service rate. As a result, in each equilibrium, the rational agents learn from the queue length upon arrival. They associate long queues with high quality.

For values of $k$ that are higher than, but close to $\Delta_{sf}$, the differentiation in service rates between the two firms is the highest ($\mu^*_l$ is about $\pi$ while $\mu^*_h = \mu$). As a result, the agent joining equilibrium has a large joining threshold.

As $k$ increases from $\Delta_{sf}$ to $\Delta_{ff}$ or $\Delta_{ss}$, the expected service rates become less differentiated and, as a result, the agent joining equilibrium threshold falls. When $k$ is lower than, but close to $\Delta_{ff}$ or $\Delta_{ss}$, the agent joining equilibrium drops to $n_{ss}$ or $n_{ff}$.

In the range of base costs of speeding up, $k \in [\Delta_{sf}, \Delta_{ss}]$, which is a subinterval of $k \in [\Delta_{sf}, \Delta_{ff}]$, a pure strategy equilibrium in service rates also exists (see Lemma 5); both firms speed up. In this equilibrium, there is no learning by rational agents from the queue length. Notice that when one firm randomizes, it is indifferent between selecting the slow or fast service rate, hence, it has the same profits as the other firm that does not randomize. For a range of base costs, the joining threshold when both firms speed up is lower than the threshold when the high quality firm is slower than the low quality firm, as discussed in Proposition 2. Hence, the ‘branch’ with profits (for both firms) corresponding with the mixed strategy equilibrium in Figure 6(B) has higher profits than
the equilibrium profits with the pure strategy. This is interesting; the service rate randomization equilibrium in which the high quality firm is slower leads to higher profits for both firms than a pure strategy service rate equilibrium in which both firms are fast. Hence, we provide a rationale for high quality firms to slow down. The latter has been a suspicion in many potential customers’ minds, but, to the best of our knowledge, has not been formally studied in a queueing game theoretic context.

When the service rate is low (case \(ii\)), a similar equilibrium in which the high quality firm is slower than the low quality firm can be characterized. The marginal revenues of speeding up are now ranked as follows: \(\Delta_{ss} < \Delta_{ff} < \Delta_{sf}\). In the range costs of speeding up, \(k \in [\Delta_{ss}, \Delta_{sf}]\), a mixed strategy equilibrium in service rates exists (see Lemma 5). On Figure 6(D), the mixed strategy equilibrium can again co-exist with a pure strategy equilibrium in which both firms are slow over \([\Delta_{ss}, \Delta_{sf}]\) and with a pure strategy equilibrium in which both firms are fast over \([\Delta_{ss}, \Delta_{ff}]\). Notice again from the Figure that the ‘branch’ corresponding to the mixed strategy equilibrium has higher profits than the equilibrium profits with the pure strategy.

Finally, notice also that when the service rate is low, an equilibrium with a slow, high quality firm occurs for relatively low capacity costs (compared to \(\Delta_{ss}\) and \(\Delta_{ff}\)), while the opposite is true when the service rate is high.

It is remarkable that even though there is no cost difference between the two firms, the agents learn from queues. The fundamental underlying reason is due to the existence of randomization strategies in service rates when the cost of speeding up are neither very high nor very low. That randomization then creates differentiation and learning opportunities for the strategic agents.

4.2. **Positively correlated product quality and cost of speeding up** (**\(v_h > v_f\) and \(k_h > k_f\)**).

Finally, we consider the case of unobservable service rates and service values, when the cost of speeding up is higher for high quality firm. This implies that, based on the cost structure, rational agents may expect longer queues for high quality firms. Indeed, we find that there will be a pure service rate equilibrium in which the high quality firm is slower than the low quality firm for
Figure 7  The equilibrium service strategy ($\beta^*$, panels (A) and (C)) and profit rates ($\Pi^*$, panels (B) and (D)) respectively as a function of $k$ for $v_h > v_\ell$ and $k_h = k + 0.01 > k_\ell = k$ (for section 4.2). Panels (A) and (B) are for high service rates Panels (C) and (D) are for low service rates compared to the potential demand arrival rate. The dashed lines are the low quality firm’s service strategy and profit rates. The thick lines indicate the pure strategy in which the high quality firm is slow, and the low quality firm is fast and is indicated by $\mu_{sf}$.

$k \in [\Delta_{sf} - \delta_k, \Delta_{sf}]$. This can be seen in Figure 7(A) and (C), where the thick lines (solid and dashed) indicate the equilibrium in which the high quality firm is slow and the low quality firm is fast. This naturally leads to agent learning through queue size.

Notice that this differentiated service rate equilibrium can co-exist with an undifferentiated pure strategy service rate equilibrium: for high service rates, the differentiated equilibrium co-exists with a pure strategy equilibrium in which both firms are fast. For low service rates, the differentiated equilibrium co-exists with a pure strategy equilibrium in which both firms are slow. Observe from
Figure 7(C) and (D), that the profit rates of both firms in this differentiated equilibrium are higher than the profit rates of the pure strategy undifferentiated equilibrium. This is again due to the strategic agent’s higher joining threshold as she learns from the queue length. Finally, the range of investment costs for which learning occurs is determined jointly by the size of the market demand and offered service rates.

The uncertainty about the service value (i.e. $v_h > v_f$) in combination with an investment cost differentiation either ‘pushes forward’ the range of investment costs where a pure strategy differentiating equilibrium exists at higher values compared to $\Delta_{ss}$ and $\Delta_{ff} - \delta_k$, when the service rate is less than the potential market arrival rate (see Figure 7(D)). The opposite occurs when the service rate is higher than the market demand ($\Lambda$): quality uncertainty ‘pushes backward’ that range to lower values compared to $\Delta_{ss}$ and $\Delta_{ff} - \delta_k$ (see Figure 7(D)).

Notice finally that because of the uncertainty about the service value, the pure strategy profits with the differentiating service rate may Pareto dominate the profits in the other pure and mixed strategy equilibria. Furthermore, for the pure strategy equilibria, the low quality firm’s profits are higher than or equal to the high quality firm’s profits. This is because of the disadvantage that the high quality firm has with respect to the service rate costs, and the incapability of communicating the quality to the market. As the revenue rates are the same, the high quality firm’s profits are (i) exactly $\delta_k$ less than the low quality firm’s profits when both speed up ($\mu_{ff}$), (ii) $k_h - k$ less when the high quality firm is slow and the low quality firm is fast ($\mu_{sf}$), and (iii) equal when both firms are slow ($\mu_{ss}$). For the mixed strategy equilibria, the profits of both firms are equal, as discussed in the previous subsection.

5. Summary, Insights and Conclusions
In this paper, we studied how rational agents learn information from a firm’s congestion level and whether a firm, anticipating such strategic agent behavior, has incentives to slow down the service rate.

We relaxed two critical assumptions that are commonly held in the queuing games literature. The first one is that the service rate is common knowledge. Instead, we assume that only the
costs associated with each service rate are common knowledge and we endogenize the service rate selection. The second assumption is that agents know the quality (utility) of the service for which they queue, and the cost of the firm to speed up. We assume that there is common uncertainty (i.e. perceived the same for all agents, but unknown) about the quality of the service and the cost of increasing the service rate, that may be positively correlated with each other. We endogenize the firm’s service rate selection and solve simultaneously the firm’s and agent’s service rate and queue joining problems. Our insights, we think, are novel. They show how the uncertainty about the service value combined with the un-observability of the service rate can lead to interesting equilibria: The high quality firm may be slower than the low quality firm in equilibrium, even if there is no cost difference for speeding up between the high and the low quality firms. This may lead to agent learning and higher profits (for both firms) than equilibria with undifferentiated service rates in which there is no agent learning. With a high quality firm selecting a slower service rate than the low quality firm, agents are enticed to join longer queues because they think that long queues are more likely for high quality firms. As more agents join than in the equilibrium with undifferentiated service rates, both firms’ demand (revenue) rate are higher. This is why the profit rate is higher when the high quality firm is slower than the low quality firm.

Our paper is an attempt to provide a rational modeling explanation for a often-quoted suspicion (that may emerge from frustrations associated with waiting) that some firms may deliberately slow down service. We show that, from a firm’s perspective, this may be rational. The fundamental reason is that queues are a natural signaling tool that firms deploy to overcome informational problems. From the perspective of a high quality firm, if quality cannot be communicated credibly, it may choose to increase the cost of the service process by e.g. emphasizing in advertisement campaigns, or investing in service technology and workforce training.

Our analysis reveals the following insights:

1. Learning among the agents is enhanced when firms differentiate through service rates. When the high and low quality firms select different service rates, agents learn from queues. When the high quality firm is fast, agents avoid long queues due to the quality uncertainty, despite the shorter
waits because they become less confident about the quality of the service. When the high quality firm is slow, agents join longer queues due to the quality uncertainty, as long queues makes them more confident that the service quality is high.

2. When the high quality service firm’s service rate is very slow compared to the low quality service firm’s service rate, the equilibrium queue joining strategy may not be a simple single threshold strategy. In addition to a balking threshold, some sputtering in the joining strategy may occur at some intermediate state, at which an arriving agent may randomize because of his indifference between joining or not joining.

3. When the service value is unknown and the investment costs are higher for the service quality, the above equilibrium selection is more pronounced. Higher investment costs for service capacity discourages the high quality firm to invest in faster service. Furthermore, due to additional signaling benefits of slow service, the high quality firm chooses slow service in equilibrium. In contrast, the low quality firm can only improve revenues through speeding up. This equilibrium that differentiates high- and low- quality firms can occur even if the investment costs are low.

**Further research:** We believe that our framework is novel and can lead to a whole new stream of research in which service rates remain unobservable. One approach would be to consider dynamic service rate changes. This approach is challenging because a subproblem of the equilibrium determination would then be a dynamic program. Besides solving simultaneously two dynamic programs (the rationality conditions of firm), the customer actions must also be rational, and consistent with the firm’s solution to the dynamic program. Furthermore, we have assumed that all agents are homogenous in their valuation of the service and are affected in the same way by the uncertainty. We believe this could be extended to heterogeneous valuations. Finally, we assumed that all potential agents have the same prior information about the service value. This assumption can also be relaxed. Two related papers, Debo et al. (2008) and Veeraghavan and Debo (2007) analyze heterogeneously informed potential agents, keeping the service rates known and fixed, without considerations of service rate decisions. We do believe that each of these extensions are worthwhile and add a significant new and unexplored dimension to the problem. We hope that our paper can
inspire researchers to disentangle the myriad of behavioral issues in queueing systems through a queuing game theoretic lens.

References


Appendix


Proof of Proposition 1: We fix \( \beta \), or, equivalently, we fix \( \mu \). We impose conditions (i) and (ii) of Definition 1 for a given joining profile \( \alpha \). We consider only strategy profiles of the following special form, parameterized by some \( n = (n_0, \alpha_0, n_1) \): \( \alpha(j) = 1 \) for \( 0 \leq j < n_0 \) and \( \alpha(n_0) = \alpha_0 \) and \( \alpha(j) = 1 \) for \( n_0 + 1 \leq j < n_1 \) and \( \alpha(j) = 0 \) for \( j \geq n_1 + 1 \). Then, we can rewrite with Lemma 1 \( l(n) \) as:

\[
l(0) = \Psi(n, \mu) \quad \text{and} \quad l(n) = l(0) \left( \frac{\mu_t}{\mu_h} \right)^n
\]

(for all \( n \) that are reached with positive probability on the long run). With Equations (5) and (6), we can rewrite conditions (i) and (ii) of Definition 1 as:

(i)-agents are rational when for each for all \( n \) that are reached with positive probability on the long run, they join when

\[
\Psi(n, \mu) \left( \frac{\mu_t}{\mu_h} \right)^n > (\leq) L(n, \mu) \Rightarrow \hat{\alpha}(n) = 1(0) \quad \text{and} \quad \Psi(n, \mu) \left( \frac{\mu_t}{\mu_h} \right)^n = L(n, \mu) \Rightarrow \hat{\alpha}(n) \in [0, 1]. \quad (13)
\]

(ii)-beliefs are consistent: The belief \( \hat{\gamma}(n) \) satisfies Bayes’ rule when:

\[
\hat{\gamma}(n) = \frac{\Psi(n, \mu) \left( \frac{\mu_t}{\mu_h} \right)^n}{\Psi(n, \mu) \left( \frac{\mu_t}{\mu_h} \right)^n + 1} \quad \text{for } n \geq 0.
\]

Hence, (1) the belief on the queue lengths that are reached with strictly positive probability on the long run, \( \hat{\gamma}(n) \) is completely specified by \( n \) and (2) for a given \( n \), the rationality conditions are completely determined. Equation (13) determines thus the equilibrium conditions \( \hat{n} \), which determine then \( \hat{\alpha} \).

Now, we introduce the variable \( \varphi > 0 \). In equilibrium, \( \hat{\varphi} \), will be equal to \( \Psi(\hat{n}, \mu) \). We replace
\( \Psi(n, \mu) \) in Equation (13) by \( \varphi \) and for any \( \varphi > 0 \), we define \( n_0 \) is the lowest queue length at which the agent balks:

\[
\hat{n}(\varphi, \mu) = \min\{n \in \mathbb{N} : \varphi \left( \frac{\mu_l}{\mu_h} \right)^n \geq L(n, \mu), \text{ for } 0 \leq n < n_0 \text{ and } \varphi \left( \frac{\mu_l}{\mu_h} \right)^n < L(n, \mu), \text{ for } n \geq n_0 \}.
\]

Recall that we introduced

\[
\Phi(\hat{n}(\varphi, \mu), \mu) = \Psi((\hat{n}(\varphi, \mu), 0, \cdot), \mu)
\]

The \( \cdot \) indicates that \( \Psi((\hat{n}_0, 0, \cdot), \mu) \) does not depend on \( n_1 \). As \( \hat{n}(\varphi, \mu) \in \mathbb{N} \), \( \hat{n}(\varphi, \mu) \) is discontinuously increasing in \( \varphi \). Hence, \( \Phi(\hat{n}(\varphi, \mu), \mu) \) is a discontinuous function in \( \varphi \). We can extend \( \Phi(\hat{n}(\varphi, \mu), \mu) \) to a correspondence, \( \hat{\Phi}(\hat{n}(\varphi, \mu), \mu) \), where at any discontinuous point \( \varphi' \) of \( \hat{n}(\varphi, \mu) \), where \( \hat{n}(\varphi - \epsilon, \mu) = n_0 \) and \( \hat{n}(\varphi + \epsilon, \mu) = n_1 \) for an arbitrary small, but strictly positive \( \epsilon \), the image of the correspondence is the set \([\Phi(n_0, \mu), \Phi(n_1, \mu)]\) if \( \Phi(n_0, \mu) < \Phi(n_1, \mu) \) or \([\Phi(n_1, \mu), \Phi(n_0, \mu)]\) otherwise.

**Existence of a fixed point of** \( \hat{\Phi}(\hat{n}(\varphi, \mu), \mu) = \varphi \). Notice that \( \varphi = 0 \): \( \hat{n}(0, \mu) \geq 0 \) for which \( \hat{\Phi}(\hat{n}(0, \mu), \mu) > 0 \) and for \( \varphi \to \infty \): \( \hat{n}(\infty, \mu) \leq \left\lfloor \frac{\mu_h}{\mu_h} \right\rfloor = \infty \) for which \( \hat{\Phi}(\hat{n}(\infty, \mu), \mu) < \infty \). It follows that the correspondence has at least one fixed point, \( \hat{\varphi} \), in \((0, +\infty)\). Next, we characterize a fixed point.

**Characterization of a fixed point of** \( \hat{\Phi}(\hat{n}(\varphi, \mu), \mu) = \varphi \). We consider three cases:

**Case 1.** Assume that for some continuous point of \( \hat{n}(\varphi, \mu) \), \( \hat{\varphi} \), the corresponding \( \hat{n}_0 = \hat{n}(\hat{\varphi}, \mu) \) satisfies

\[
\hat{\varphi} = \Phi(\hat{n}(\hat{\varphi}, \mu), \mu)
\]

Then, it is easy to see that the strategy profile \( \hat{\alpha} \), defined by \( \hat{\alpha}(j) = 1 \) for \( 0 \leq j < \hat{n}_0 \) and \( \hat{\alpha}(\hat{n}_0) = 0 \) satisfies (i) of Definition 1, by construction of \( \hat{n}(\varphi, \mu) \) (as \( \hat{n}_0 \) is the lowest queue length at which the agent balks, the agent joins for all queue lengths strictly lower than \( \hat{n}_0 \), as is assumed in the special structure of \( \alpha \)) and also (ii) of Definition 1 because \( \hat{\varphi} = l(0) \). Hence, \( \hat{\alpha} \) satisfies conditions (i) and (ii) of Definition 1.

**Case 2.** Assume that no continuous point of \( \hat{n}(\varphi, \mu) \) exists for which the above equality is satisfied.
As the correspondence $\hat{\Phi}(\hat{n}(\hat{\varphi}, \bm{\mu}), \bm{\mu})$, there must exist at least one discontinuous point, $\hat{\varphi}$, that satisfies $\hat{\Phi}(\hat{n}(\hat{\varphi}, \bm{\mu}), \bm{\mu}) = \hat{\varphi}$. Now, let $\hat{n}_0 = \hat{n}(\hat{\varphi} - \epsilon, \bm{\mu})$ and assume that $\hat{n}(\hat{\varphi} + \epsilon, \bm{\mu}) = \hat{n}_0 + 1$ then:

$$\Phi(\hat{n}_0, \bm{\mu}) > \hat{\varphi} > \Phi(\hat{n}_0 + 1, \bm{\mu}) \text{ or } \Phi(\hat{n}_0, \bm{\mu}) < \hat{\varphi} < \Phi(\hat{n}_0 + 1, \bm{\mu})$$  \hspace{1cm} (14)

Now consider $\alpha(j) = 1$ for $0 \leq j < n_0$ and $\alpha(n_0) = \alpha_0$ and $\alpha(j) = 0$ for $j \geq n_0 + 1$. Then, condition (ii) for consistent beliefs becomes:

$$\hat{\varphi} = \Psi((\hat{n}_0 + 1, \alpha_0, \hat{n}_0 + 1), \bm{\mu})$$

From continuity of the right hand side of the above expression in $\alpha_0$, which ranges from $\Phi(\hat{n}_0, \bm{\mu})$ (for $\alpha_0 = 0$) to $\Phi(\hat{n}_0 + 1, \bm{\mu})$ (for $\alpha_0 = 1$) and by the inequalities (14) it follows that there must exist a $\hat{\alpha}_0$ such that the rationality condition at the empty queue needs to be satisfied: $\hat{\varphi} = \Psi((\hat{n}_0 + 1, \alpha_0, \hat{n}_0 + 1), \bm{\mu})$. Then, it is easy to see that the strategy profile $\hat{\alpha}$, defined by $\hat{\alpha}(j) = 1$ for $0 \leq j < \hat{n}_0$, $\hat{\alpha}(\hat{n}_0) = \hat{\alpha}_0$ and $\hat{\alpha}(j) = 0$ for $j \geq \hat{n}_0 + 1$ satisfies (i) of Definition 1, by construction of $\hat{n}(\varphi, \bm{\mu})$ and also (ii) of Definition 1 because $\hat{\varphi} = l(0)$. Hence, $\hat{\alpha}$ satisfies conditions (i) and (ii) of Definition 1.

**Case 3.** Assume that for some $\hat{\varphi}$, the conditions of (2.) cannot be satisfied for $\hat{n}(\hat{\varphi} + \epsilon, \bm{\mu}) = \hat{n}_0 + 1$, for some $\hat{n}_1 > \hat{n}_0 + 1$. Then, either:

$$\Phi(\hat{n}_0, \bm{\mu}) > \hat{\varphi} > \Phi(\hat{n}_1, \bm{\mu}) \text{ or } \Phi(\hat{n}_0, \bm{\mu}) < \hat{\varphi} < \Phi(\hat{n}_1, \bm{\mu}).$$  \hspace{1cm} (15)

Similarly as in condition 2., consider $\alpha(j) = 1$ for $0 \leq j < n_0$ and $\alpha(n_0) = \alpha_0$ and $\alpha(j) = 1$ for $n_0 + 1 \leq j < n_1$ and $\alpha(j) = 0$ for $j \geq n_1 + 1$. Notice that by definition of $\Psi$ (Equation 7) for $\hat{n}_1 = \hat{n}(\varphi + \epsilon, \bm{\mu}) > \hat{n}_0 + 1$:

$$\Phi(\hat{n}_0, \bm{\mu}) = \Psi((\hat{n}_0, 0, \hat{n}_1), \bm{\mu}) \text{ and } \Phi(\hat{n}_1, \bm{\mu}) = \Psi((\hat{n}_0, 1, \hat{n}_1), \bm{\mu}).$$

From continuity of $\Psi((\hat{n}_0, \alpha_0, \hat{n}_1), \bm{\mu})$ in $\alpha_0$, and by the inequalities (15) it follows that there must exist a $\hat{\alpha}_0$ such that the rationality condition at the empty queue needs to be satisfied: $\hat{\varphi} = \Psi((\hat{n}_0, \hat{\alpha}_0, \hat{n}_1), \bm{\mu})$.\hspace{1cm}
\[ \Psi ((\hat{n}_0, \hat{\alpha}_0, \hat{n}_1), \mu). \] Again, it is easy to see that the strategy profile \( \hat{\alpha} \), defined by \( \hat{\alpha}(j) = 1 \) for \( 0 \leq j < n_0 \) and \( \hat{\alpha}(n_0) = \hat{\alpha}_0 \) and \( \hat{\alpha}(j) = 1 \) for \( n_0 + 1 \leq j < n_1 \) and \( \hat{\alpha}(j) = 0 \) for \( j \geq n_1 + 1 \) satisfies (i) of Definition 1 (by construction of \( \hat{n}(\varphi, \mu) \)) and (ii) of Definition 1 (because \( \hat{\varphi} = \ell(0) \)). Hence, \( \hat{\alpha} \) satisfies conditions (i) and (ii) of Definition 1.

Cases 1., 2. and 3. result in Equations (7), (8), (9) and (10).

**Proof of Lemma 2:** Define:

\[ \varphi(\nu, \mu) = \begin{cases} \left( \frac{\nu}{\mu} \right) L(\nu, \mu) & \text{for } \nu \in [\nu_+ (\mu), \nu_- (\mu) - 1) \\ +\infty & \text{for } \nu \geq \nu_- (\mu) - 1, \end{cases} \]

Where \( \nu_+ (\mu) = v_+ \mu_c / c \). \( \varphi(\nu, \mu) \) is the likelihood ratio upon observing an empty queue that will make the agent indifferent between joining a queue of length \( n \) and balking, assuming that at queue lengths 0 to \( n \), all agents join and informed agents only join when the firm is of high quality. I.e. when \( \hat{n}(\varphi', \mu) = n' \), then \( \varphi(n', \mu) = \varphi' \). When \( \varphi(n, \mu) \) is increasing in \( n \), this means that as \( \varphi \) increases, \( \hat{n}(\varphi, \mu) \) increases in increments of 1. When \( \varphi(n, \mu) \) is not monotone, this means that \( \hat{n}(\varphi, \mu) \) increases in increments of potentially more than 1. In order to establish this, we consider \( \nu \) as a continuous variable and derive \( \varphi(\nu) \). We find that over \( [\nu_+ (\mu), \nu_- (\mu) - 1) \), the derivative can have zero or two zero points. We ignore the knife-edge case of one degenerate zero point. Let \( \nu_+ (\mu) \) and \( \nu_- (\mu) \) be the zero points of \( \frac{d}{d\nu} \varphi(\nu, \mu) \). It can easily be established that \( \frac{d}{d\nu} \varphi(0, \mu) > 0 \), hence, \( \varphi(\nu, \mu) \) increases first, reaches a local maximum and then decreases to a local minimum, after which it increases again. It can also be established easily that \( \lim_{\nu \to \nu_- (\mu) - 1} \varphi(\nu, \mu) = +\infty \). As a result, over \( [\nu(\mu), \nu_- (\mu)], \) where \( \varphi(\nu, \mu) \) decreases, \( \hat{n}(\varphi, \mu) \) may increase in jumps that are larger than 1. Now, we take the derivative:

\[
\frac{d}{d\nu} \varphi(\nu, \mu) = \left\{ \frac{v_+ - v_-}{(v_+ - \frac{c^{\nu+1}}{\mu}})^2 c \frac{v_+ - c^{\nu+1}}{\mu v_+ - c^{\nu+1}} \ln \left( \frac{\mu h}{\mu c} \right) \left[ \frac{1}{l_0} \left( \frac{\mu h}{\mu c} \right)^\nu \right] \right\}
\]

Hence, it follows that

\[
\frac{d}{d\nu} \varphi(\nu, \mu) > 0 \iff \frac{c^{\nu+1}}{\mu v_+ - c^{\nu+1}} \left( v_+ - \frac{c^{\nu+1}}{\mu v_+ - c^{\nu+1}} \right) + \frac{c^{\nu+1}}{\mu v_+ - c^{\nu+1}} \left( \ln \left( \frac{\mu h}{\mu c} \right) \right) > 0
\]
\[ \frac{c}{\mu_\ell} + \frac{(-v_\ell + c \frac{\nu + 1}{\mu_\ell})}{\mu_h} + \frac{(-v_\ell + c \frac{\nu + 1}{\mu_\ell})}{\mu_h} \ln \left( \frac{\mu_h}{\mu_\ell} \right) > 0 \iff \]

\[ c \frac{\mu_h v_h - \mu_\ell v_\ell}{\mu_h \mu_\ell} + \left( c \frac{\nu + 1}{\mu_\ell} - v_\ell \right) \left( v_h - c \frac{\nu + 1}{\mu_h} \right) \ln \left( \frac{\mu_h}{\mu_\ell} \right) > 0 \]

\[ \Rightarrow -\ln \left( \frac{\mu_h}{\mu_\ell} \right) \left( \frac{c \frac{\nu + 1}{\mu_\ell} - v_\ell}{v_h - c \frac{\nu + 1}{\mu_h}} \right) \geq 0 \]

This is a quadratic equation in \( \nu \). Hence, \( \frac{d}{d\nu} \varphi(\nu, \mu) = 0 \) has zero, or two real solutions, depending on the parameters. Notice that the left hand side reaches a maximum when

\[ \frac{d}{d\nu} \left( -v_\ell + c \frac{\nu + 1}{\mu_\ell} \left( v_h - c \frac{\nu + 1}{\mu_h} \right) \right) = 0 \iff \nu^* = -1 + \frac{\mu_h v_h + \mu_\ell v_\ell}{2c} \]

\[ v_\ell - c \frac{\nu^* + 1}{\mu_\ell} < 0 < v_h - c \frac{\nu^* + 1}{\mu_h} \iff \frac{1}{2} \frac{\mu_h v_h - \mu_\ell v_\ell}{\mu_\ell} < 0 < \frac{1}{2} \frac{\mu_h v_h - \mu_\ell v_\ell}{\mu_h} \]

and as a result, assuming that \( \mu_h v_h - \mu_\ell v_\ell > 0 \) is always satisfied because in the worst case, \( \frac{\mu_h}{\mu_\ell} > \frac{\nu + 1}{\mu_\ell} \) (all other cases are trivially satisfied), \( \nu^* \) lies in between \( \nu_\ell(\mu) \) and \( \nu_h(\mu) - 1 \). Thus, we can write:

\[ \left( -v_\ell + c \frac{\nu + 1}{\mu_\ell} \right) \left( v_h - c \frac{\nu + 1}{\mu_h} \right) \leq \frac{1}{4} \left( \frac{\mu_h v_h - \mu_\ell v_\ell}{\mu_\ell} \right)^2 \]

Plugging Equation (17) in Equation (16), we obtain that:

\[ c \frac{\mu_h v_h - \mu_\ell v_\ell}{\mu_h \mu_\ell} > -\frac{1}{4} \left( \frac{\mu_h v_h - \mu_\ell v_\ell}{\mu_\ell} \right)^2 \ln \left( \frac{\mu_h}{\mu_\ell} \right) \Rightarrow \frac{d}{d\nu} \varphi(\nu, \mu) > 0, \forall \nu \in [\nu_\ell, \nu_h - 1] \]

or

\[ C(\mu) = c + \frac{1}{4} \left( \mu_h v_h - \mu_\ell v_\ell \right) \ln \left( \frac{\mu_h}{\mu_\ell} \right) > 0 \Rightarrow \frac{d}{d\nu} \varphi(\nu, \mu) > 0, \forall \nu \in [\nu_\ell, \nu_h - 1] \].

We have thus proven that when \( C(\mu) > 0 \), \( \varphi(\nu, \mu) \) is increasing. Hence \( \hat{n}(\varphi, \mu) \) increases in jumps of one. When \( C(\mu) < 0 \), there exist two roots of Equation (16), \( \nu_\ell(\mu) \) and \( \nu_h(\mu) \). \( \varphi(\nu, \mu) \) is decreasing in between the two roots, which implies that \( \hat{n}(\varphi, \mu) \) will never take values over \( (\nu_h(\mu), \nu_h(\mu)) \).

Hence \( \hat{n}(\varphi, \mu) \) may increase in jumps that are larger than one. Notice that when \( \frac{\mu_h}{\mu_\ell} > 1 \), \( C(\mu) > 0 \).

Only when \( \frac{\mu_h}{\mu_\ell} < 1 \), we may obtain that \( C(\mu) < 0 \).

**Proof of Lemma 3:** First, we write \( \Phi(\nu, \mu) \) as follows:

\[ \Phi(\nu, \mu) = \frac{\sum_{n=0}^{\nu} \left[ \frac{\Lambda}{\bar{\nu}_\ell} \right]^n}{\sum_{n=0}^{\nu} \left[ \frac{\Lambda}{\bar{\nu}_h} \right]^n} = \frac{1 - \left( \frac{\Lambda}{\bar{\nu}_\ell} \right)^{\nu + 1}}{1 - \left( \frac{\Lambda}{\bar{\nu}_h} \right)^{\nu + 1}} \]
We show that $\Phi(\nu, \mu)$ increases in $\nu$ iff $\mu_\ell < \mu_h$. Now, we consider $\nu$ as a continuous variable and determine the condition when $\Phi(\nu)$ is increasing:

$$
\frac{d}{d\nu} \Phi(\nu, \mu) > 0 \iff \frac{1}{1 - \left[\frac{\Lambda}{\mu_h}\right]^{\nu+1}} > \Phi(\nu, \mu) \iff
\frac{1 - \frac{\Lambda}{\mu_h}}{1 - \left[\frac{\Lambda}{\mu_h}\right]^{\nu+1}} > \Phi(\nu, \mu)
$$

as $\frac{d}{d\nu} (a^{\nu+1}) = a^{\nu+1} \ln a$, we obtain that

$$
\frac{d}{d\nu} \Phi(\nu, \mu) > 0 \iff \left(\frac{\mu_h}{\mu_\ell}\right)^{\nu+1} \frac{1 - \frac{\Lambda}{\mu_h} \ln \left(\frac{\Lambda}{\mu_h}\right)}{1 - \frac{\Lambda}{\mu_\ell} \ln \left(\frac{\Lambda}{\mu_\ell}\right)} > \frac{1 - \frac{\Lambda}{\mu_h}}{1 - \left[\frac{\Lambda}{\mu_h}\right]^{\nu+1}}
$$

and $\frac{d}{d\nu} \Phi(\nu, \mu) > 0 \iff \left(\frac{\Lambda}{\mu_\ell}\right)^{\nu+1} \frac{1 - \frac{\Lambda}{\mu_\ell} \ln \left(\frac{\Lambda}{\mu_\ell}\right)}{1 - \left[\frac{\Lambda}{\mu_\ell}\right]^{\nu+1}} > \frac{1 - \frac{\Lambda}{\mu_h}}{1 - \left[\frac{\Lambda}{\mu_h}\right]^{\nu+1}}$

(18)

Now, consider

$$
\nu^{\nu+1} \ln (x) \frac{1}{1 - x^{\nu+1}} \text{ then}
$$

and let $z = x^{\nu+1}$, then notice

$$
\frac{\ln \left(\frac{z}{1 - z}\right)}{1 - z} = \frac{1}{\nu + 1} \frac{\ln (z)}{1 - z}
$$

and as $\frac{\ln (z)}{1 - z} < 0$ and continuous decreasing for $z > 0$. Hence, the condition of Equation (18) is satisfied iff $\frac{\Lambda}{\mu_\ell} > \frac{\Lambda}{\mu_h}$ or $\mu_h > \mu_\ell$, then $\frac{d}{d\nu} \Phi(\nu, \mu) > 0$, otherwise $\frac{d}{d\nu} \Phi(\nu, \mu) < 0$.

**Proof of Lemma 4:** The result follows immediately from the threshold strategy determined by $[v\mu/c]$, where $\mu$ is the appropriate service rate.

**Proof of Proposition 2:**

**Part (i): Comparative static for $\hat{n}_{sf}$:** We consider pure strategy equilibria. We prove the comparative static for $\mu_{sf} = (\mu_h, \mu)$ in $\mu_h$: $\hat{n}_{sf}$ is decreasing in $\mu_h = \frac{\mu}{\mu} = \frac{\mu_h}{\mu_\ell}$ for $\mu_h \in (\mu_\ell, \bar{\mu})$ when $\bar{\mu}$ is sufficiently high. Let $\hat{n}_{ff}$ be the pure strategy equilibrium for $\mu_{ff} = (\bar{\mu}, \bar{\mu})$. For notational convenience, we drop from now onwards the subscript $sf$ from $\mu_{sf}$. Recall that $\Phi(n, \mu) = \Psi((n, 0, \bullet), \mu)$. 

Consider a point of discontinuity, \( \varphi(n, \mu) \), at any \( n \) that satisfies \( c(n + 1)/\mu_h < v_h \): \( \hat{n}(\hat{\varphi} - \epsilon, \mu) = n \) and \( \hat{n}(\hat{\varphi} + \epsilon, \mu) = n + 1 \). In other words:

\[
\varphi(n, \mu) = \left( \frac{\mu_h}{\mu_l} \right)^n L(n, \mu).
\]

A pure strategy equilibrium characterized by a threshold, \( n = \hat{n}_{sf} \), satisfies

\[
\varphi(n - 1, \mu) < \Phi(n, \mu) < \varphi(n, \mu).
\]

where

\[
\Phi(n, \mu) = \frac{1 - (\frac{\mu}{\mu_h})^{n+1}}{1 - (\frac{\mu}{\mu_h})} \quad \text{and} \quad \varphi(n, \mu) = \left( \frac{\mu_h}{\mu} \right)^n \frac{1 - v_h + c(n + 1) \frac{1}{\mu}}{l_0 v_h - c(n + 1) \frac{1}{\mu_h}}.
\]

In order to obtain the comparative statics of the pure strategy equilibrium in \( \mu_h \), we need to ‘track’ a horizontal line segment with value \( \Phi(\hat{n}_{ff}, \mu) \) for \( \varphi \in (\varphi(\hat{n}_{ff} - 1, \mu), \varphi(\hat{n}_{ff}, \mu)) \) when \( \mu_h \) decreases from \( \mu \) to \( \mu_o \), i.e., we want to find the locus of the end points of the segment; 

\[ \{ \varphi(\hat{n}_{ff} - 1, \mu), \varphi(\hat{n}_{ff}, \mu) \} \quad \text{and} \quad \{ \varphi(\hat{n}_{ff} + 1, \mu), \varphi(\hat{n}_{ff}, \mu) \} \]

for \( \mu_h \in (\mu_o, \mu] \). For a given \( \mu \), these loci can be represented as functions \( \Phi(\varphi) \) and \( \Phi(\varphi) \). This is illustrated in Figure 8. We show that the horizontal line segment \( \{ \varphi, \Phi(\hat{n}_{ff}, \mu) \} \) for \( \varphi(\hat{n}_{ff} - 1, \mu) < \varphi < \varphi(\hat{n}_{ff} + 1, \mu) \) moves to the left and down and crosses the 45 degree line at most once as \( \mu_h \) decreases from \( \mu \) to \( \mu_o \). This implies that a pure strategy equilibrium cannot be lower than \( \hat{n}_{sf} \), which proves the comparative static for \( \hat{n}_{sf} \).

The end points of the horizontal segment are: \( \{ \varphi(\hat{n}_{ff} - 1, \mu), \Phi(\hat{n}_{ff}, \mu) \} \) and \( \{ \varphi(\hat{n}_{ff} + 1, \mu), \Phi(\hat{n}_{ff}, \mu) \} \) for \( \mu_h \in (\mu_o, \mu] \). Where these loci are functions, the are denoted as \( \Phi(\varphi) \) and \( \Phi(\varphi) \) respectively. The first derivative of the end points is:

\[
\frac{\partial}{\partial \varphi} \Phi(\varphi) = \frac{\partial}{\partial \mu_h} \Phi(\hat{n}_{ff}, \mu) \quad \text{and} \quad \frac{\partial}{\partial \varphi} \Phi(\varphi) = \frac{\partial}{\partial \mu_h} \Phi(\hat{n}_{ff}, \mu)
\]

We can obtain that:

\[
\frac{\partial}{\partial \mu_h} \varphi(\hat{n}_{ff}, \mu) = \frac{1}{\mu_h} \left[ \frac{\mu_h v_h}{c} \varphi(n, \mu) \right]
\]
Figure 8 Illustration of segment corresponding with $n_{ff} = 5$ for four values of $\mu_h$: 1.250, 1.173, 0.939, 0.906, 0.890, when $\mu_l = \bar{\mu} = 1.25$. Other parameters are: $c = 0.151$, $p = 0.35$, $r = 1$, $\Lambda = 1$, $v_h = 1.5$, $v_l = 0.25$. The plotted lines are $[\varphi(n - 1, \mu), \Phi(n, \mu)]$ and $[\varphi(n, \mu), \Phi(n, \mu)]$.

and

$$\frac{\partial}{\partial \mu_h} \Phi(n_{ff}, \mu_h) = \frac{1}{\mu_h} F(\frac{\Lambda}{\mu_h}, n_{ff}) \Phi(n_{ff}, \mu_h)$$

where

$$F(\rho, n) = \frac{\rho}{1 - \rho^{n+1}} \frac{1 - (n+1)\rho^n + n\rho^{n+1}}{1 - \rho}$$

and

$$G(v, n) = \frac{vn - (n+1)^2}{v - (n+1)}$$

Sign of $F(\rho, n)$: As (i) $\lim_{\rho \to 1} F(\rho, n) = \frac{1}{2} n > 0$, (ii) we can rewrite:

$$\frac{1 - (n+1)\rho^n + n\rho^{n+1}}{1 - \rho} = \sum_{k=0}^{n-1} \rho^k - n\rho^n,$$

which, as it only contains one sign change in the coefficients of $\rho^k$, it must have at most one root (due to Newton’s rule of signs), which is 1, and is positive for $\rho \in (0, 1)$ and negative for $\rho \in (1, +\infty)$, while $\frac{\rho}{1 - \rho^{n+1}}$ is positive for $\rho \in (0, 1)$ and negative for $\rho \in (1, +\infty)$, hence the product is always positive for $\rho \in (0, 1) \cup (1, +\infty)$. As a result of (i) and (ii), $F(\rho, n)$ is positive for $\rho > 0$ for any $n$. 
Sign of $G(v,n)$: It is easy to see that $G(v,n) > 0$ for $v > (n+1)^2 / n$ (which is larger than $n+1$).

We have obtained that

$$\frac{\partial}{\partial \phi} \Phi(\phi) = \frac{F(\Lambda_{\hat{n}_{ff} \mu}, \hat{n}_{ff})}{G(\frac{\Lambda_{\hat{n}_{ff} \mu}}{c}, \hat{n}_{ff})} \frac{\partial}{\partial \phi} \Phi(\hat{n}_{ff}, \mu)$$

and letting

$$\mu_h > \mu_a = \frac{c}{v_h} (\hat{n}_{ff} + 1)^2 \Rightarrow \frac{\partial}{\partial \phi} \Phi(\phi) > 0$$

Now, define $\mu_b$ as:

$$\mu_b = \min \{ \mu : F(\Lambda_{\mu}, \hat{n}_{ff}) < G(\frac{\Lambda_{\mu}}{c}, \hat{n}_{ff}) \}$$

then, consider any $\mu_h \geq \max\{ \mu_a, \mu_b \}$. There are two possibilities:

- $\overline{\Phi}(\hat{n}_{ff}, \mu) > \varphi(\hat{n}_{ff}, \mu)$: Then $\frac{\partial}{\partial \phi} \overline{\Phi}(\phi) > 0$, i.e. $\overline{\Phi}(\phi)$ increases. As by definition of $\hat{n}_{ff}$, for $\mu_h = \overline{\mu}$, $\overline{\Phi}(\hat{n}_{ff}, (\mu, \overline{\mu})) < \varphi(\hat{n}_{ff}, (\mu, \overline{\mu}))$, there must exist by continuity of $\overline{\Phi}(\phi)$ a value of $\mu'$ in $(\max\{ \mu_a, \mu_b \}, \overline{\mu})$ such that $\Phi(\hat{n}_{ff}, (\mu', \overline{\mu})) = \varphi(\hat{n}_{ff}, (\mu', \overline{\mu}))$. For $\mu_h \in (\max\{ \mu_a, \mu_b \}, \overline{\mu})$ as both factors of $\frac{\partial}{\partial \phi} \overline{\Phi}(\phi)$ are less than 1, we obtain that $0 < \frac{\partial}{\partial \phi} \overline{\Phi}(\phi) < 1$. Hence, $\overline{\Phi}(\phi)$ cannot intersect the 45 degree line over $(\max\{ \mu_a, \mu_b \}, \overline{\mu})$ and $\overline{\Phi}(\phi)$ crosses thus exactly once.

- $\overline{\Phi}(\hat{n}_{ff}, \mu) < \varphi(\hat{n}_{ff}, \mu)$: Then, for $\max\{ \mu_a, \mu_b \} < \mu_h < \overline{\mu}$, $0 < \frac{\partial}{\partial \phi} \overline{\Phi}(\phi) < 1$, hence, $\overline{\Phi}(\phi)$ cannot intersect the 45 degree line and does not cross the 45 degree line.

The analysis for $\Phi(\phi)$ is similar, except that as $\Phi(\hat{n}_{ff}, \mu) > \varphi(\hat{n}_{ff} - 1, \mu)$, it is impossible that there is a crossing with the 45 degree line when

$$\mu_h \geq \max\{ \mu'_a, \mu'_b \}$$

where

$$\mu'_a = \frac{c}{v_h} \hat{n}_{ff}^2 \text{ and } \mu'_b = \min \{ \mu_h : F(\frac{\Lambda}{\mu}, \hat{n}_{ff}) < G(\frac{\Lambda_{\mu h}}{c}, \hat{n}_{ff} - 1) \}$$

Note that $\mu'_a < \mu_a$. Hence, we obtain that for $\mu_h \geq \mu_a \equiv \max\{ \mu_a, \mu'_a, \mu_b \}$, the end points of the horizontal segment $\{ [\varphi(\hat{n}_{ff} - 1, \mu), \Phi(\hat{n}_{ff}, \mu)] \}$ moves downward to the left and crosses the 45 degree line at most once as $\mu_h$ decreases from $\overline{\mu}$ to $\mu_a$. As a result, the equilibrium $\hat{n}_{sf}$ cannot be lower than $\hat{n}_{ff}$.
Part (ii): Comparative static for $\hat{n}_{fs}$: The proof is analogous and illustrated in Figure 9. We summarize: We prove the comparative static for $\mu_{fs} = (\mu_h, \mu)$ in $\mu_h$: $\hat{n}_{fs}$ is decreasing in $\mu_h = \bar{\mu}$ for $\mu_h \in [\mu, +\infty)$. Let $\hat{n}_{ss}$ be the pure strategy equilibrium for $\mu_{ss} = (\bar{\mu}, \bar{\mu})$. We show that the horizontal line segment $[\varphi, \Phi(\hat{n}_{ss}, \mu)]$ for $\varphi(\hat{n}_{ss} - 1, \mu) < \varphi < \varphi(\hat{n}_{ss}, \mu)$ moves to the right and up and crosses the 45 degree line at most once as $\mu_h$ decreases from $\underline{\mu}$ to $+\infty$. Hence, the horizontal segment, $\{[\varphi(\hat{n}_{ss} - 1, \mu), \Phi(\hat{n}_{ss}, \mu)], [\varphi(\hat{n}_{ss}, \mu), \Phi(\hat{n}_{ss}, \mu)]\}$ must intersect with the 45 degree line. This implies that a pure strategy equilibrium must be smaller than or equal to $\hat{n}_{ss}$, which proves the comparative static for $\hat{n}_{ss}$. Similarly as in Part (i), the condition for $0 < \frac{\partial}{\partial \varphi} \Phi(\varphi) < 1$ are

$$\mu_a = \frac{c}{v_h} \left( \hat{n}_{ss} + 1 \right)^2$$

and

$$\mu_b = \min \{ \mu_h : F \left( \frac{\Lambda}{\mu_h}, \hat{n}_{ss} \right) < G \left( \frac{\mu_h v_h}{c}, \hat{n}_{ss} \right) \}$$

and

$$\mu'_b = \min \{ \mu : F \left( \frac{\Lambda}{\mu}, \hat{n}_{ff} \right) < G \left( \frac{\mu v_h}{c}, \hat{n}_{ff} - 1 \right) \}$$
Proof of Lemma 5: It can be seen that conditions (i) to (iii) are all immediate consequence of Definition 1(iii) that \( \mu^* \) is a pure strategy equilibrium service rate.

Conditions (iv) and (v) are proven as follows. Consider any \( k \) such that:

- \( \Delta_{sf} < k < \Delta_{ss} \): Now, define \( \mu(\beta) = (\mu_\beta (\beta/\mu + (1 - \beta)/\mu)^{-1}) \), from which: \( \mu(0) = \mu_{ss} \) and \( \mu(1) = \mu_{sf} \). With Proposition 1, there exist at least one pure strategy or mixed strategy threshold agent equilibrium for every \( \beta \in [0,1] \). Denote this agent equilibrium with \( \hat{n}(\beta), \hat{\alpha}(\beta) \). It is easy to see that the firm’s revenues result in:

\[
\Delta(\hat{n}(\beta), \hat{\alpha}(\beta)) \text{ decreases from } \Delta(\hat{n}_{ss}, \hat{\alpha}_{ss}) = \Delta_{ss} \text{ to } \Delta(\hat{n}_{sf}, \hat{\alpha}_{sf}) = \Delta_{sf}
\]

Now, we claim that any \( \Delta \in [\Delta_{sf}, \Delta_{ss}] \) will be reached by some \( \beta \in [0,1] \).

1. When \( \hat{\alpha}(\beta) > 0 \), the characterization of the mixed strategy for the agents is a function of \( \beta \) (determined by the indifference condition of the agents at \( \hat{n}(\beta) \)). As long as \( \hat{\alpha}(\beta) > 0 \), \( \hat{n}(\beta) \) does not change. Let us denote \( \hat{n}(\beta) \) by \( \hat{n} \). Hence, the firm’s revenues are a continuous function of \( \beta \). It is easy to see that \( \hat{\alpha}(\beta) \) is monotone and hence also the firm’s profit rate, hence, also \( \Delta(\hat{n}, \hat{\alpha}(\beta)) \).

2. When \( \hat{\alpha}(\beta) = 0 \), the characterization of the pure strategy for the agents is not a function of \( \beta \). As long as \( \hat{\alpha}(\beta) = 0 \), \( \hat{n}(\beta) \) does not change. Let us denote \( \hat{n}(\beta) \) again by \( \hat{n} \). Hence, the firm’s revenues are independent of \( \beta \). Hence \( \Delta(\hat{n}, 0) \) is a constant as long as \( \hat{\alpha}(\beta) = 0 \).

Since a randomization strategy exists for any \( \Delta \in [\Delta_{sf}, \Delta_{ss}] \) that leads to \( \Delta \) marginal revenues for any \( k \in [\Delta_{sf}, \Delta_{ss}] \), there exists a \( \beta \) such that the marginal revenues are \( \Delta \), making the high quality firm indifferent between \( \mu \) and \( \mu \), such that condition (iii) of Definition 1 is satisfied.

- \( \Delta_{sf} - k_h - k_f < k < \Delta_{ff} - k_h - k_f \): Now, define \( \mu(\beta) = (\beta/\mu + (1 - \beta)/\mu)^{-1} \), from which: \( \mu(0) = \mu_{sf} \) and \( \mu(1) = \mu_{ff} \). The reasoning is similar as for the first bullet: For any \( \Delta \in [\Delta_{sf} - k_h - k_f < k < \Delta_{ff} - k_h - k_f] \), a mixing probability \( \beta \) can be found that results in marginal revenues \( \Delta \), making the high quality firm indifferent between \( \mu \) and \( \mu \), such that condition (iii) of Definition 1 is satisfied.

Proof of Proposition 3: Suppose \( \mu \) is close enough to \( \mu \) such that \( R(n, \mu) \approx R(n, \mu) + \frac{\partial R(n, \mu)}{\partial \mu}(\mu - \mu) \), when
\[
\Delta = R(n, \mu) - R(n, \mu) \approx \frac{\partial R(n, \mu)}{\partial \mu} (\mu - \mu)
\]

We show that
\[
\frac{\partial R(n, \mu)}{\partial \mu} < \frac{\partial R(n + 1, \mu)}{\partial \mu} \text{ for } \mu < \Lambda \text{ and }
\frac{\partial R(n, \mu)}{\partial \mu} > \frac{\partial R(n + 1, \mu)}{\partial \mu} \text{ for } \mu > \Lambda.
\]

Let \( \rho = \Lambda / \mu \) and assume that \( \rho \neq 1 \), then:
\[
r_n(\rho) = \frac{d}{d \rho} \sum_{k=0}^{n-1} \rho^k = \frac{-n\rho^{n-1} + \rho^n - \rho^{2n} + n\rho^n}{(1 - \rho^{n+1})^2}
\]

In Step 1, we prove that \( r_n(\rho) - r_{n+1}(\rho) \) has exactly one root in \((0, +\infty)\). We call that root \( \rho_n \).

In Step 2, we show that \( \rho_n \in (0, 1) \). In Step 3, we show that \( \rho_n \) increases in \( n \). These three steps prove Cases i and ii of the Proposition.

**Step 1.** Now, we prove that \( r_n(\rho) - r_{n+1}(\rho) \) has exactly one root in \((0, +\infty)\). We obtain
\[
\frac{-n\rho^{n-1} + \rho^n - \rho^{2n} + n\rho^n}{(1 - \rho^{n+1})^2} = -\frac{\rho^n(1 - \rho)^3}{\rho(1 - \rho^{n+1})^2 (1 - \rho^{n+2})^2} \left( \sum_{k=0}^{2n+2} n\rho^k - \sum_{k=0}^{2n+2} \rho^k \min(2k, 4n + 3 - 2k) \right)
\]

where
\[
P(\rho, n) = \left( \sum_{k=0}^{2n+2} n\rho^k - \sum_{k=0}^{2n+2} \rho^k \min(2k, 4n + 3 - 2k) \right) \quad (19)
\]

Notice that \( r_n(1) = \lim_{\rho \to 1} \frac{-n\rho^{n-1} + \rho^n - \rho^{2n} + n\rho^n}{(1 - \rho^{n+1})^2} = -\frac{n}{2^{n+1}} \) and hence \( r_n(1) - r_{n+1}(1) = -\frac{n}{2^{n+1}} + \frac{1}{2^{n+2}} \geq 0 \). So, \( \rho = 1 \) can never be a root of \( r_n(\rho) - r_{n+1}(\rho) \). Hence, the roots of \( r_n(\rho) - r_{n+1}(\rho) \) are equal to the roots of \( P(\rho, n) \) that are not equal to 1.

It follows from the above expression that the coefficients on the RHS of Equation (19) changes signs two times. As a result, \( P(\rho, n) \) has at most two roots over \((0, +\infty)\) (Newton’s rule of signs).

It can easily be seen that for \( \rho = 1 \), the RHS of equation (19) becomes
\[
\sum_{k=0}^{2n+1} (n - \min(2k, 4n + 3 - 2k)) + n + 1 = 0
\]
(due to symmetry) hence, \( \rho = 1 \) is one root of \( P(\rho, n) \), then, there is another one at most one other root, \( \rho_n \), of \( r_n(\rho) - r_{n+1}(\rho) \).

**Step 2.** We now show that \( \rho_n \in (0, 1) \): Note that \( k < 2n - k + \frac{3}{2} \iff k < n + \frac{3}{4} \). Hence,

\[
P(\rho, n) = \sum_{k=0}^{2n+2} n\rho^k - \sum_{k=0}^{n} 2k\rho^k - \sum_{k=n+1}^{2n+2} 2(2n - k + \frac{3}{2})\rho^k
\]

\[= \sum_{k=0}^{n} n\rho^k - \sum_{k=0}^{n} 2k\rho^k + \sum_{k=0}^{n+1} n\rho^{n+1+k} - \sum_{k=0}^{n+1} (2k - 1)\rho^{2n+2-k} \]

\[= \sum_{k=0}^{n} (n-2k)\rho^k + \rho^{n+1} \sum_{k=0}^{n+1} ((n+1)\rho^k - 2k\rho^{n+1-k}) + \rho^{n+1} \sum_{k=0}^{n+1} (\rho^{n+1-k} - \rho^k) \]

\[= \sum_{k=0}^{n} (n-2k)\rho^k + \rho^{n+1} \sum_{k=0}^{n+1} ((n+1)\rho^k - 2k\rho^{n+1-k}) \]

Note \( \sum_{k=0}^{n+1} ((n+1)\rho^k - 2k\rho^{n+1-k}) \) has at most one root in \((0, +\infty)\) as there is only one sign change in the coefficients. The root is \( \rho = 1 \) as \( \sum_{k=0}^{n+1} (n+1) = 0 \). It is easy to show that the term is strictly positive for \( \rho > 1 \), hence:

\[
P(\rho, n) = \sum_{k=0}^{n} (n-2k)\rho^k + \rho^{n+1} \sum_{k=0}^{n+1} ((n+1)\rho^k - 2k\rho^{n+1-k}) \]

\[\geq \sum_{k=0}^{n} (n-2k)\rho^k + \sum_{k=0}^{n+1} ((n+1)\rho^k - 2k\rho^{n+1-k}) \]

\[= -\sum_{k=0}^{n} \rho^k + (n+1)\rho^{n+1} \]

It is easy to see again that \(-\sum_{k=0}^{n} \rho^k + (n+1)\rho^{n+1} \) has at most one root in \((0, +\infty)\) as there is only one sign change in the coefficients. The root is \( \rho = 1 \) as \(-\sum_{k=0}^{n} 1 + (n+1) = 0 \). As a result:

For \( \rho > 1 \): \( P(\rho, n) > 0 \).

Hence, the second root of \( P(\rho, n) \) must be in \((0, 1)\).

**Step 3.** We now show that \( \rho_n \) increases in \( n \): We can rewrite the equation that determines \( \rho_n \), \( P(\rho, n) = 0 \) as follows:
\[ 1 = \rho^{-n-1} \sum_{k=0}^{n} (n-2k) \rho^k \]

\[ \sum_{k=0}^{n+1} (- (n+1) \rho^k + 2k \rho^{n+1-k}) \]

It is easy to see that for \( \rho \in (0,1) \) the numerator and denominator of right hand side are positive.

We will show that:

\[ \rho^{-n-1} \frac{\sum_{k=0}^{n+1} (n-2k) \rho^k}{\sum_{k=0}^{n+1} (- (n+1) \rho^k + 2k \rho^{n+1-k})} \text{ increases in } n \text{ for any } \rho \in (0,1) \]

Hence, \( \rho_n \) increases in \( n \). We rewrite the difference

\[ \rho^{-n-2} \frac{\sum_{k=0}^{n+2} (n+1-2k) \rho^k}{\sum_{k=0}^{n+2} (- (n+2) \rho^k + 2k \rho^{n+2-k})} - \rho^{-n-1} \frac{\sum_{k=0}^{n} (n-2k) \rho^k}{\sum_{k=0}^{n+1} (- (n+1) \rho^k + 2k \rho^{n+1-k})} \]

as

\[ \left( \sum_{k=0}^{n+1} (n+1-2k) \rho^k \right) \left( \sum_{k=0}^{n+1} (- (n+1) \rho^k + 2k \rho^{n+1-k}) \right) \]

\[ \left( \sum_{k=0}^{n+2} (- (n+2) \rho^k + 2k \rho^{n+2-k}) \right) \left( \sum_{k=0}^{n} (n-2k) \rho^k \right) \]

\[ +(1-\rho) \sum_{k=0}^{n} (n-2k) \rho^k \sum_{k=0}^{n+2} (- (n+2) \rho^k + 2k \rho^{n+2-k}) \]

divided by \( \left( \sum_{k=0}^{n+2} (- (n+2) \rho^k + 2k \rho^{n+2-k}) \right) \times \left( \sum_{k=0}^{n+1} (- (n+1) \rho^k + 2k \rho^{n+1-k}) \right) \). As the latter is always positive, the sign of Equation (20) is determined by the expression above. The third term is positive for \( \rho \in (0,1) \). The first two terms can be rewritten as \( \frac{T(\rho,n)}{(1-\rho)\rho^{n+1}} \) where \( T(\rho,n) = 1 + \rho^{2n+4} - (n+2)^2 \rho^{n+1} \left( \rho^2 - 2 \frac{(n+3)(n+1)}{(n+2)^2} \rho + 1 \right) \). Note that

\[ \frac{d^2}{d\rho^2} \left( \frac{1 + \rho^{2n+4}}{\rho^{n+1}} \right) = \frac{(n+2)(n+1+3\rho^{2n+4}+n\rho^{2n+4})}{\rho^{n+3}} > 0 \]

and the second degree Taylor series expansion at \( \rho = 1 \) gives:

\[ \frac{1 + \rho^{2n+4}}{\rho^{n+1}} = (n+2)^2 \left( \rho^2 - 2 \frac{(n+3)(n+1)}{(n+2)^2} \rho + 1 \right) + O ((1-\rho)^3) \]
Hence, it follows that

\[ \frac{1 + \rho^{2n+4}}{\rho^{n+1}} \geq (n + 2)^2 \left( \rho^2 - 2(n + 3) (n + 1) \frac{\rho}{(n + 2)^2} + 1 \right) \]

and we obtain that \( T(\rho, n) \geq 0 \) and hence, \( \rho_n \) increases in \( n \).