



Manufacturing & Service Operations Management

Publication details, including instructions for authors and subscription information:
<http://pubsonline.informs.org>

Signaling Service Quality Through Queue Disclosure

Pengfei Guo, Moshe Haviv, Zhenwei Luo, Yulan Wang

To cite this article:

Pengfei Guo, Moshe Haviv, Zhenwei Luo, Yulan Wang (2023) Signaling Service Quality Through Queue Disclosure. Manufacturing & Service Operations Management 25(2):543-562. <https://doi.org/10.1287/msom.2022.1170>

Full terms and conditions of use: <https://pubsonline.informs.org/Publications/Librarians-Portal/PubsOnLine-Terms-and-Conditions>

This article may be used only for the purposes of research, teaching, and/or private study. Commercial use or systematic downloading (by robots or other automatic processes) is prohibited without explicit Publisher approval, unless otherwise noted. For more information, contact permissions@informs.org.

The Publisher does not warrant or guarantee the article's accuracy, completeness, merchantability, fitness for a particular purpose, or non-infringement. Descriptions of, or references to, products or publications, or inclusion of an advertisement in this article, neither constitutes nor implies a guarantee, endorsement, or support of claims made of that product, publication, or service.

Copyright © 2022 The Author(s)

Please scroll down for article—it is on subsequent pages



With 12,500 members from nearly 90 countries, INFORMS is the largest international association of operations research (O.R.) and analytics professionals and students. INFORMS provides unique networking and learning opportunities for individual professionals, and organizations of all types and sizes, to better understand and use O.R. and analytics tools and methods to transform strategic visions and achieve better outcomes.

For more information on INFORMS, its publications, membership, or meetings visit <http://www.informs.org>

Signaling Service Quality Through Queue Disclosure

Pengfei Guo,^a Moshe Haviv,^{b,c} Zhenwei Luo,^{d,*} Yulan Wang^d

^a College of Business, City University of Hong Kong, Kowloon, Hong Kong; ^b School of Data Science, Shenzhen Research Institute of Big Data, The Chinese University of Hong Kong, Shenzhen Campus, China; ^c Department of Statistics and Data Science and the Federmann Center for the Study of Rationality, The Hebrew University of Jerusalem, Jerusalem 91904, Israel; ^d Faculty of Business, The Hong Kong Polytechnic University, Kowloon, Hong Kong

*Corresponding author

Contact: penguo@cityu.edu.hk,  <https://orcid.org/0000-0002-0447-3556> (PG); moshe.haviv@gmail.com (MH); zhen-wei.luo@polyu.edu.hk,  <https://orcid.org/0000-0002-4446-4363> (ZL); yulan.wang@polyu.edu.hk,  <https://orcid.org/0000-0003-4184-590X> (YW)

Received: September 14, 2020

Revised: October 24, 2021; June 6, 2022;
October 13, 2022


Accepted: October 24, 2022

Published Online in Articles in Advance:
December 23, 2022

<https://doi.org/10.1287/msom.2022.1170>

Copyright: © 2022 The Author(s)

Abstract. *Problem definition:* We consider a single-server queueing system where service quality is either high or low. The server, who knows its exact quality level, can signal this quality information to customers by revealing or concealing its queue length. Based on this queue disclosure action and the observed queue length in the case of a revealed queue, customers decide whether to join the system. *Academic/practical relevance:* The queue disclosure action is regarded as a signal indicating the service quality. *Methodology:* We develop a signaling game and adopt the sequential equilibrium concept to solve it. We further apply the perfect sequential equilibrium as an equilibrium-refinement criterion. *Results:* In our baseline model, where all of the customers are uninformed of service quality, the pure-strategy perfect sequential equilibrium is always a pooling one, except at several discrete values of market size (measured by the potential arrival rate). When the market size is below a certain threshold, both high- and low-quality servers adopt queue concealment; otherwise, both types of servers adopt queue revelation. We also consider a general scenario in which the market is composed of both quality informed and uninformed customers. Under this setting, when the server conceals the queue, we can fully characterize customers' equilibrium queueing strategies and the corresponding effective arrival rates. The unique sequential equilibrium outcome is still a pooling one when the market size is either below a lower threshold or above an upper threshold. A separating equilibrium can occur only when the market size falls between two thresholds; under that circumstance, the uninformed customers can infer the server's quality from its queue disclosure behavior. *Managerial implications:* Under separating sequential equilibria, uninformed customers can fully infer the quality information and thus behave in an informed way. Unlike studies where queue disclosure is not regarded as a quality signal, our study reveals that the signaling effect of queue disclosure increases (decreases) the effective arrival rate of the high-quality (low-quality) server and also increases the customers' total utility when the server is of low quality.

 **Open Access Statement:** This work is licensed under a Creative Commons Attribution- NonCommercial-NoDerivatives 4.0 International License. You are free to download this work and share with others, but cannot change in any way or use commercially without permission, and you must attribute this work as "Manufacturing & Service Operations Management. Copyright © 2022 The Author(s). <https://doi.org/10.1287/msom.2022.1170>, used under a Creative Commons Attribution License: <https://creativecommons.org/licenses/by-nc-nd/4.0/>."

Funding: P. Guo acknowledges the financial support from the Research Grants Council of Hong Kong [Grant 15502820]. The research of M. Haviv was funded by Israel Science Foundation [Grant 1512/19]. Z. Luo acknowledges the financial support from the Internal Start-up Fund of the Hong Kong Polytechnic University [Grant P0039035] and the National Natural Science Foundation of China [Grant 71971184]. Y. Wang's work was supported by the Research Grants Council of Hong Kong [Grant 15505019].

Supplemental Material: The e-companion is available at <https://doi.org/10.1287/msom.2022.1170>.

Keywords: queueing theory • service quality • signaling game • queue disclosure • sequential equilibrium

1. Introduction

In many service systems, the service quality of a server is known to some customers but unknown to others. Take a restaurant as an example. The food quality depends on factors such as the chefs' skills and the quality of the ingredients. Local residents are likely to know the quality of the food in a restaurant, whereas tourists might

not. Customers who know the service quality are called *informed customers*, and those who do not are called *uninformed customers*. According to Debo et al. (2012), uninformed customers can infer a server's quality by inspecting the queue length. That study finds that the queueing behavior of uninformed customers follows a "hole-avoiding" strategy: they behave almost the same

as customers informed of high quality, except that when the queue reaches a certain queue length, called the hole, they do not join. Their work assumes that the queue is always observable. Here, we consider a scenario in which the server can control the visibility of its queue. Several questions arise from the situation in which the server has the choice of keeping the queue visible or concealing it. Why is a server willing to reveal its queue length to customers? Isn't the visibility of the queue length itself a signal indicating service quality?

In this paper, we investigate these questions by treating the visibility of the queue length as a signal of service quality. Unlike the scenario in Debo et al. (2012), where the uninformed customers are able to infer quality only from the behavior of the informed customers, here, the uninformed customers also infer quality based on the server's queue disclosure behavior. With advanced information technologies, the visibility of queues can now be easily adjusted in many settings, particularly in virtual/online queueing systems. For example, call centers can choose to either play music or inform waiting customers of the number of customers waiting in front of them. Restaurants can choose to reveal or hide the queue length information when customers order foods on mobile apps or online platforms such as Dianping.com or DoorDash.com. Even physical restaurants can show the number of waiting customers on display screens or queueing machines (e.g., KFC and McDonald's have installed such machines). Indeed, advances in information technology have made it easy for servers to convey queue information to customers. The question is whether the server has an incentive to reveal this information, particularly when taking the quality of the service into account.

In this paper, we consider a setting where the queue disclosure is a decision of the server. In this scenario, an uninformed customer can infer service quality information not only by inspecting the queue length (if the queue is observable) but also from the server's queue disclosure action. Specifically, we consider the following signaling game setting between a server and customers. The customers arrive to a single-server queueing system according to a Poisson process. Their service times follow an exponential distribution. The server's service quality is either high or low, which is determined by nature via a Bernoulli trial. A fraction of the customers are informed, and they know the server's exact quality level. They are considered to be *positively informed* if the server is of high quality or *negatively informed* if the server is of low quality. The uninformed customers have prior beliefs about the server's service quality. The customers are otherwise identical; that is, they receive the same service reward and bear the same per-unit-time delay cost. The server knows its own quality and can choose to either reveal or conceal the queue length. The queue disclosure action signals the server's service quality to customers. The server's goal is to maximize the effective arrival

rate. Upon observing the server's queue disclosure action and the actual queue length (if the queue length is revealed), the uninformed customers update their beliefs about the server's quality type and then decide whether to join the queue.

To solve this signaling game, we apply the *sequential equilibrium* solution concept (Kreps and Wilson 1982). Hereafter, we simply call it equilibrium. Given the existence of multiple equilibria, we adopt the *perfect sequential equilibrium* (Grossman and Perry 1986) as the refinement criterion. Three types of equilibria are considered: the *pure-strategy equilibrium* in which both high- and low-quality servers choose to either always reveal or always conceal the queue, the *mixed-strategy equilibrium* in which both types of servers randomize revealing and concealing the queue, and the *hybrid-strategy equilibrium* in which one type of server chooses to either always reveal or always conceal the queue and the other type randomizes these two actions. Furthermore, for the pure-strategy equilibrium, when both types of servers adopt the same action, we call it a *pooling equilibrium*, but when each type takes a different action, we call it a *separating equilibrium*.

We start with a benchmark scenario where the system only has uninformed customers. If the signaling effect of the queue disclosure action is not considered, the uninformed customers will treat the service quality level as the expected level, and their joining strategy then follows the classic strategy stated in Hassin and Haviv (2003). In a signaling setting, if the high- and low-quality servers adopt different queue disclosure strategies, then the customers can infer the service quality level from the server's disclosure action. Clearly, the high-quality server attracts more customers. This creates an incentive for the low-quality server to mimic the disclosure behavior of the high-quality server. When there are only uninformed customers in the system, we demonstrate that mimicking is the best choice for the low-quality server. It can help the low-quality server to attract more customers, because they cannot tell the server's quality level. Specifically, except at some discrete values of market size (i.e., the potential arrival rate), the unique pure-strategy perfect sequential equilibrium is that both types of servers choose to conceal (reveal) the queue when the market size is below (above) a threshold. The pooling equilibrium cannot convey quality information to customers. Therefore, the effective arrival rates of the high- and low-quality servers remain the same as those under the *nonsignaling case*, in which the queue disclosure action is not regarded as a quality signal.

The above conclusion, however, does not hold in a setting with heterogeneous customers, that is, when some customers are informed but others are uninformed. The informed customers do not need to rely on a signal to judge the server's quality, and their queueing behavior contains some quality information. The uninformed customers can then infer the server's quality from the observed

queue length, as demonstrated in Debo et al. (2012). These make it more difficult for the low-quality server to mimic the high-quality server, and thus a separating equilibrium can be sustained. Indeed, we show that in a very small (large) market, both types of servers choose to conceal (reveal) their queues in equilibrium, but when the market size falls into an intermediate range, the equilibrium outcome may be a separating one, with one type of server concealing the queue and the other type revealing the queue. The existence of a separating equilibrium yields three new insights into our signaling mechanism. First, it brings new understanding to the queueing behavior of uninformed customers. Recall that in our setting, uninformed customers have three sources of information about service quality: (1) the prior belief, (2) the signal of the queue disclosure action, and (3) the queue length if it is observable. Our results show that in small- and large-sized markets, source 2 does not provide useful information, and thus uninformed customers rely on sources 1 and 3 to make their joining or balking decision. When the queue is observable, the uninformed customers adopt a hole-avoiding joining strategy, as demonstrated in Debo et al. (2012). By contrast, in a medium-sized market where a separating equilibrium exists, source 2 provides the full information about the server's quality type, and thus source 3 becomes redundant. In this situation, an uninformed customer behaves exactly the same as an informed one. These results enrich the conclusion of Debo et al. (2012), who only considered information sources 1 and 3 for uninformed customers in an observable queue. Second, we show that when separating sequential equilibria exist, the maximum effective arrival rate of the high-quality (low-quality) server considering all pure-strategy perfect sequential equilibria is larger (smaller) than the optimal one when the queue disclosure action is not considered a signal. In other words, signaling benefits the high-quality server but harms the low-quality server. Third, customers benefit from such a signaling mechanism when the server is of low quality, because it helps them to avoid an overly crowded queueing system. However, the signaling mechanism may not be able to improve the customers' total utility when the server is of high quality: although it can help customers to fully infer the server's high quality, too many customers might be encouraged to join, which may not be socially desirable. We further find that as the proportion of informed customers increases, the range of market sizes at which separating equilibria can be sustained does not necessarily expand.

In practice, it is common for service providers to communicate their service quality through advertising or offering free trials, which is costly. Our study provides a nice insight for them: it is unnecessary for a service provider to set a target of "educating" all customers. As long as a fraction of customers are informed, the remaining

uninformed ones can infer the service quality from the service provider's queue disclosure action, which is often costless. However, one needs to be cautious that such a strategy only works in a medium-sized market.

The remainder of this paper is organized as follows. We review the related literature in Section 2. In Section 3, we present the signaling game and introduce the definitions of sequential equilibrium and perfect sequential equilibrium. The sequential equilibrium of a baseline model in which all of the customers are uninformed is analyzed in Section 4. The equilibrium analyses for the general scenario with heterogeneous customers are presented in Section 5. We further conduct the sensitivity analysis to examine the impact of customer type composition and service price on the existence of separating sequential equilibria in Section 6. Concluding remarks are provided in Section 7. We relegate the supplementary materials to Online Appendix A, and all of the proofs are provided in Online Appendix B.

2. Literature Review

The research considering strategic customers in queueing systems originates from Naor (1969). In this research stream, our work is related to the studies on delay announcements. Hassin (1986) investigated a server's incentive to disclose the queue length information and found that the server prefers concealing (revealing) the queue in a small (large) market. Other studies that have investigated the impact of delay announcements include Whitt (1999), Armony and Maglaras (2004a, 2004b), Burnetas and Economou (2007), Guo and Zipkin (2007), Armony et al. (2009), Guo and Hassin (2011), Yu et al. (2016), Ibrahim et al. (2017), Yu et al. (2017), Hu et al. (2018), and Yu et al. (2021). We refer interested readers to the two survey books, Hassin and Haviv (2003) and Hassin (2016), and the review papers of Aksin et al. (2007) and Ibrahim (2018) and references therein for studies in this research stream.

Our study is also related to the stream of research on information provision and purchase in queues. Hassin and Haviv (1994) examined a parallel queueing system in which customers can buy information on the queue lengths to join the shorter queue. Hassin (2007) studied a scenario in which the server knows its service quality and other system parameters and decides whether to disclose such information. In Hassin and Roet-Green (2017), customers may balk, join directly, or buy the queue length information first and then make their joining or balking decisions. Hassin and Roet-Green (2018) further considered a setting with parallel servers in which an uninformed customer becomes informed after paying to inspect the queues. In our study, we do not consider information purchase. Instead, the uninformed customers can infer the server's quality level from the server's queue disclosure action.

Our study is closely related to the literature on signaling games in queueing systems. Allon et al. (2011) considered a *cheap talk* game in which a server sends a queue-length-dependent signal to customers. Yu et al. (2018) studied a cheap talk game with heterogeneous customers and showed that the server can infer customer types through customers' reaction to the server's delay announcement. Veeraraghavan and Debo (2009, 2011) considered two parallel queues in which uninformed customers could infer some quality information from observable queue lengths. Debo et al. (2012) considered an observable queue with both informed and uninformed customers. They showed that uninformed customers' pure equilibrium joining strategy is a hole-avoiding joining strategy. Many recent studies have considered other quality signals, such as service or waiting times (Debo and Veeraraghavan 2014, Kremer and Debo 2016), price and wait lines (Debo et al. 2020), and information generated by customers (Yu et al. 2016, Wang and Hu 2020). Unlike these studies, here we consider a signaling game where the server's queue disclosure action is a signal of its quality level.

In a signaling game, the sender takes a signaling action after the realization of the state of the world. We note that some recent studies of queueing systems consider a different timing sequence of the game by adopting Bayesian persuasion (Kamenica and Gentzkow, 2011), under which the sender pre-commits to a strategy before the state of the world is realized. Lingenbrink and Iyer (2019) applied Bayesian persuasion to a queueing setting by considering that the server pre-commits to a queue-length-dependent signaling strategy. Guo et al. (2022) showed that under the uncertain service quality, the server can ex ante commit to a quality-dependent queue disclosure strategy before the service quality is realized to persuade more customers to join the system. Different from these studies, we consider a case in which the server has no commitment power and makes the queue disclosure decision after the realization of quality type.

3. Model Setup

In this section, we describe our signaling game and introduce the concept of the sequential equilibrium and the associated equilibrium-refinement criterion.

3.1. Timing of the Signaling Game

Consider a single-server queueing system. Nature moves first and determines the server's quality type t according to a Bernoulli distribution: with probability δ , the server is of high quality (labeled H) and with probability $1 - \delta$, it is of low quality (labeled L), where $0 < \delta < 1$. After observing its own quality type t ($t \in \mathbb{T} := \{H, L\}$), the server chooses a queue disclosure action, that is, revealing the queue (denoted by R) or concealing the queue (denoted by C) as a signal to convey its quality information to

customers. Let $\mathbb{S} := \{R, C\}$ denote the server's signal set. The customers arrive at the server according to a Poisson process with rate λ , with prior knowledge of service quality being high with probability δ . Service times are independent and identically distributed exponential random variables with rate μ . Define $\rho := \lambda/\mu$. All of the customers who join the system receive the same quality of service and incur the same waiting cost of θ per unit time in the system (waiting time plus service time). When a customer is served by a high-quality (low-quality) server, the customer receives a monetary reward V_H (V_L). The inequality $V_H > V_L > \frac{\theta}{\mu}$ is required to ensure that at least one customer joins the system. Upon observing the server's revealing (concealing) action, the uninformed customers update their belief of the server being the high-quality type to δ^R (δ^C), and correspondingly, the belief of the server being the low-quality type is updated to $1 - \delta^R$ ($1 - \delta^C$). The customers then decide whether to join the system. When the server conceals the queue, the queue is unobservable (labeled U). In this case, the customer's queueing strategy can be represented by the probability that the customer will join the system. When the server chooses to reveal the queue, the queue becomes observable (labeled O). Let $\pi_{i,H}(\delta^R)$ ($\pi_{i,L}(\delta^R)$) denote the steady-state probability of the queue length being i ($i = 0, 1, \dots$) for the high-quality (low-quality) server. Upon observing the queue length i , the uninformed customers further update their beliefs (i.e., the probability that the server is of high quality), denoted by $Pr(H | i, \delta^R)$, and then make their joining or balking decisions. We normalize the server's reward for serving a customer to 1. Then, the server's payoff can be measured as equivalent to the customers' effective arrival rate.

The timing of our signaling game is summarized as follows.

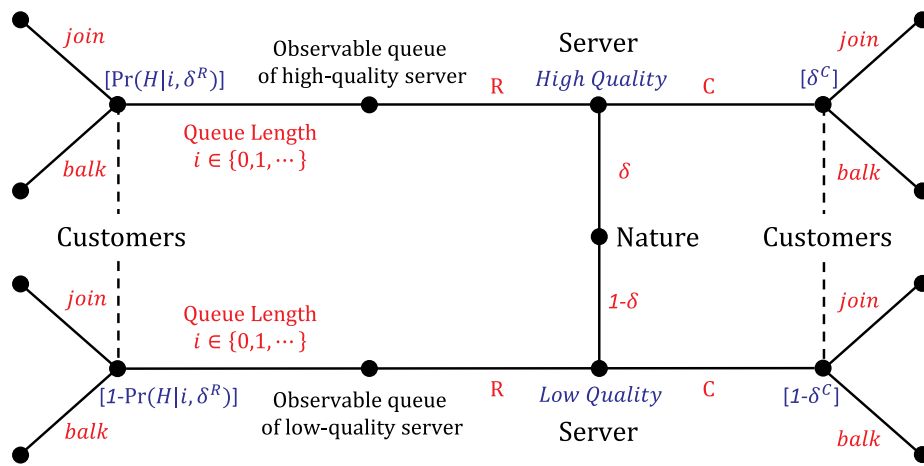
- (1) Nature chooses the server's service quality type t ($t \in \mathbb{T} = \{H, L\}$) according to a Bernoulli distribution.
- (2) The server learns its quality type and then chooses a queue disclosure action from the set $\mathbb{S} = \{R, C\}$.
- (3) The uninformed customers update their beliefs about the server's service quality based on the queue disclosure action and, if the queue is observable, the queue length.
- (4) The customers make their joining or balking decisions.

This signaling game can be represented in an extensive form, as shown in Figure 1.

3.2. Definitions of the Sequential Equilibrium and Perfect Sequential Equilibrium

In this study, we apply the *sequential equilibrium* concept (Kreps and Wilson 1982) to solve our signaling game, which is defined as follows.

Figure 1. (Color online) Extensive Form of the Signaling Game (With Payoffs Ignored for Simplicity)



Definition 1 (Sequential Equilibrium). A sequential equilibrium of the signaling game is a behavior-belief profile consisting of the server’s signaling rules $f(s | t)$, where $f(s | t)$ specifies the probability that the type t ($t \in \mathbb{T}$) server chooses signal s ($s \in \mathbb{S}$), the customers’ joining rules, and the customers’ beliefs δ^C , δ^R , and $Pr(H | i, \delta^R)$ ($i = 0, 1, \dots$), which should satisfy the following two conditions:

(i) (Sequential Rationality) No player deviates from the equilibrium strategy at each of its information sets under the specified beliefs.

(ii) (Consistency) When the server sends signal $s \in \mathbb{S}$ with a positive probability, the customers update their beliefs using signal s according to the Bayes’ rule; that is, if $\delta f(s | H) + (1 - \delta)f(s | L) > 0$, then $\delta^s = \frac{\delta f(s | H)}{\delta f(s | H) + (1 - \delta)f(s | L)}$. After observing queue length i ($i = 0, 1, \dots$) in a revealed queue, the uninformed customers further update their beliefs as $Pr(H | i, \delta^R) = \frac{\delta^R \pi_{i,H}(\delta^R)}{\delta^R \pi_{i,H}(\delta^R) + (1 - \delta^R) \pi_{i,L}(\delta^R)}$.

In an equilibrium, if the server sends a signal with a positive probability, we say that this signal is *on the equilibrium path*; otherwise, it is *off the equilibrium path*. Condition (ii) in the above definition does not put any restriction on the customers’ off-equilibrium-path posterior belief. This may lead to multiple equilibria, and some of them may be unreasonable. Here, we adopt the *perfect sequential equilibrium* (Grossman and Perry 1986) as a further refinement criterion to impose restrictions on the off-equilibrium-path beliefs. In addition to the above two conditions, the perfect sequential equilibrium essentially requires the following *credibility (of the updating rule)* for our signaling game.

Definition 2 (Credible Updating Rule). For a signal s that is off the equilibrium path, given the customers’ equilibrium queueing strategies under the signal s and their new belief (satisfying the credible updating rule), denote the set of types of servers that can be

strictly better off by deviating from the equilibrium strategy to s by \mathbb{T}' and the set of types of servers that are indifferent between deviating to s and staying at the equilibrium strategy by \mathbb{T}'' . Let $h(t)$ be the probability of type t ($t \in \mathbb{T}$) server deviating from the equilibrium strategy to s , which should satisfy $h(t) = 1$ if $t \in \mathbb{T}'$, $h(t) \in [0, 1]$ if $t \in \mathbb{T}''$ and $h(t) = 0$ if $t \in \mathbb{T} / (\mathbb{T}' \cup \mathbb{T}'')$. If there exists a nonempty set $\mathbb{T}' \cup \mathbb{T}''$, then

(a) the customers’ posterior belief about type t upon observing the signal s is $\frac{w(t)h(t)}{\sum_{t' \in \mathbb{T}' \cup \mathbb{T}''} w(t')h(t')}$ (we require $\sum_{t \in \mathbb{T}' \cup \mathbb{T}''} h(t) > 0$), where $w(\cdot)$ denotes the prior belief (i.e., $w(H) = \delta$ and $w(L) = 1 - \delta$),¹ and

(b) the sets \mathbb{T}' and \mathbb{T}'' remain unchanged under the above posterior belief. Otherwise, there is no restriction on customers’ posterior belief after seeing the off-equilibrium-path signal s .

Under Definition 2, when the uninformed customers’ posterior belief γ satisfies the credible updating rule given an off-equilibrium-path signal, the corresponding equilibrium queueing strategy will make the types of servers in the set \mathbb{T}' strictly better off if they deviate with probability 1 and the types of servers in the set \mathbb{T}'' indifferent between remaining on the equilibrium path and deviating with some probability such that the customers’ posterior belief is γ . The process of identifying a belief that satisfies the credible updating rule is essentially a fixed-point argument.

In our queueing setting, when both types of servers strictly prefer to deviate, the credible updating rule requires the customers’ off-equilibrium-path posterior belief to be equal to the prior; when only the high-quality (low-quality) server strictly prefers to deviate or is indifferent between deviating and remaining on the equilibrium path but the low-quality (high-quality) server strictly prefers to remain on the equilibrium path, it requires the off-equilibrium-path posterior belief to be 1 (0). These help to eliminate some unreasonable sequential equilibrium outcomes and

narrow down the range of the off-equilibrium-path belief. For the case in which both types of servers strictly prefer to remain on the equilibrium path (i.e., the set $\mathbb{T}' \cup \mathbb{T}''$ is empty), the credible updating rule puts no restrictions on the off-equilibrium-path belief.

4. Equilibrium Analysis

In this section, we investigate a benchmark scenario in which all of the customers are uninformed. We first derive the customers' equilibrium queueing strategies in the cases of observable and unobservable queues in Section 4.1. We then conduct the equilibrium analysis for the whole signaling game in Section 4.2. In Section 4.3, we examine the signaling effect on the effective arrival rate.

4.1. Customers' Equilibrium Queueing Strategies and Effective Arrival Rates

Because all of the customers are identical, we consider only symmetric strategies for the customers in the queueing game. If the server reveals the queue, the uninformed customers update their belief as $Pr(H | i, \delta^R) = \frac{\delta^R \pi_{i,H}(\delta^R)}{\delta^R \pi_{i,H}(\delta^R) + (1 - \delta^R) \pi_{i,L}(\delta^R)}$ based on observed queue length i ($i = 0, 1, \dots$). Because all of the customers are uninformed, the probabilities for the queue length being i are the same for both the high-quality and low-quality servers (i.e., $\pi_{i,H}(\delta^R) = \pi_{i,L}(\delta^R)$). Hence, we can obtain the posterior $Pr(H | i, \delta^R) = \delta^R$ for all queue lengths i . In other words, the queue length contains no information about service quality when all the customers are uninformed. Based on Naor (1969), we know that the uninformed customers all join if and only if the queue length (including the one in service) upon arrival does not exceed the threshold $n(\delta^R) := \lfloor [\delta^R V_H + (1 - \delta^R) V_L] \mu / \theta \rfloor - 1$, where $\lfloor \cdot \rfloor$ is the floor function. In the following analysis, we simply use $n(\delta^R)$ to denote the customers' equilibrium queueing strategy in a revealed queue. In steady state, the system under a revealed queue is an $M/M/1/(n(\delta^R) + 1)$ queue with a capacity constraint of $n(\delta^R) + 1$. Let $p_{n(\delta^R)+1}$ be the probability that the queue length is $n(\delta^R) + 1$. Then, $p_{n(\delta^R)+1} = \rho^{n(\delta^R)+1} / \sum_{i=0}^{n(\delta^R)+1} \rho^i$. Denote the customers' effective arrival rate to this observable queue as $\lambda^O(\delta^R)$. We then have

$$\lambda^O(\delta^R) = \lambda(1 - p_{n(\delta^R)+1}) = \frac{\lambda \sum_{k=0}^{n(\delta^R)} \rho^k}{\sum_{i=0}^{n(\delta^R)+1} \rho^i}. \quad (1)$$

Note that we can rewrite $\lambda^O(\delta^R) = \mu - \frac{\mu}{\sum_{i=0}^{n(\delta^R)+1} \rho^i}$. Clearly, $\lambda^O(\delta^R)$ is strictly increasing in the potential arrival rate λ .

We then consider the case where the server conceals the queue and the uninformed customers' belief is δ^C . According to Edelson and Hildebrand (1975), if the

potential arrival rate λ is small enough ($\lambda < \mu - \theta / [\delta^C V_H + (1 - \delta^C) V_L]$), all of the customers join the system; otherwise, the customers in equilibrium adopt a mixed strategy; that is, they join the queue with probability $\frac{\mu - \theta / [\delta^C V_H + (1 - \delta^C) V_L]}{\lambda}$. Denote the customers' equilibrium joining probability as $p(\delta^C)$. We then have

$$p(\delta^C) = \begin{cases} 1, & \text{if } \lambda < \mu - \theta / [\delta^C V_H + (1 - \delta^C) V_L]; \\ \frac{\mu - \theta / [\delta^C V_H + (1 - \delta^C) V_L]}{\lambda}, & \text{otherwise.} \end{cases} \quad (2)$$

Let $\lambda^U(\delta^C)$ be the customers' effective arrival rate to this unobservable queue. It can be derived that

$$\lambda^U(\delta^C) = \begin{cases} \lambda, & \text{if } \lambda < \mu - \theta / [\delta^C V_H + (1 - \delta^C) V_L]; \\ \mu - \theta / [\delta^C V_H + (1 - \delta^C) V_L], & \text{otherwise.} \end{cases} \quad (3)$$

4.2. Sequential Equilibrium Analysis

We are now ready to derive the server's equilibrium signaling strategy. When all of the customers are uninformed, the sequential rationality condition in Definition 1 is equivalent to the following requirements: the customers' joining rule is $(n(\delta^R), p(\delta^C))$, and the server's signaling rule maximizes its expected payoff such that $\forall t \in \mathbb{T}, f(R | t) > 0$ ($f(C | t) > 0$) only if $\lambda^O(\delta^R) \geq \lambda^U(\delta^C)$ ($\lambda^U(\delta^C) \geq \lambda^O(\delta^R)$). We can express a sequential equilibrium as

$$[(f(R | H), f(R | L)), (n(\delta^R), p(\delta^C)), \delta^R, \delta^C].$$

There are three types of signaling strategies: a *pure strategy* in which both high- and low-quality servers choose to either always reveal or always conceal the queue (i.e., both $f(R | H)$ and $f(R | L)$ are either 0 or 1), a *mixed strategy* in which both types of servers randomize revealing and concealing the queue (i.e., both $f(R | H)$ and $f(R | L)$ are strictly between 0 and 1), and a *hybrid strategy* in which one type of server chooses to either always reveal or always conceal the queue and the other type randomizes these two actions (i.e., one of the two probabilities $f(R | H)$ and $f(R | L)$ is either 0 or 1, but the other one is strictly between 0 and 1). The pure strategies can be further classified into two types: the *pooling strategy* in which both types of servers send the same signal and the *separating strategy* in which the two types of servers send different signals. Here, we focus mainly on the pure strategy; the analysis on hybrid and mixed strategies can be found in the Online Appendix A.

We investigate each of the four pure strategies in turn. For simplicity, let (s', s'') with $s', s'' \in \{R, C\}$ denote the pure strategy played by two types of servers under which the high-quality server always chooses signal s' and the low-quality server always chooses s'' , that is, $f(s' | H) = 1$ and $f(s'' | L) = 1$.

(1) (R,R) , that is, *pooling on R*. Then, R is on the equilibrium path, and by Bayes' rule, the customers' updated belief after observing R is still $\delta^R = \delta$. Hence, the payoffs to both types of servers are the same and have the value $\lambda^O(\delta)$. To check whether both types of servers are willing to stay on R , we need to check the off-equilibrium-path belief δ^C . The sequential equilibrium concept does not put any restriction on δ^C . As long as the off-equilibrium-path belief δ^C satisfies $\lambda^U(\delta^C) \leq \lambda^O(\delta)$, both types of servers have no incentive to deviate to C . Therefore, the corresponding pooling equilibrium is $[(R,R), (n(\delta^R), p(\delta^C)), \delta^R = \delta, \delta^C]$, with δ^C satisfying $\lambda^U(\delta^C) \leq \lambda^O(\delta)$.

(2) (C,C) , that is, *Pooling on C*. Similarly, we can show that the profile $[(C,C), (n(\delta^R), p(\delta^C)), \delta^R, \delta^C = \delta]$ with the off-equilibrium-path belief δ^R satisfying $\lambda^O(\delta^R) \leq \lambda^U(\delta)$ is a pooling equilibrium.

(3) (R,C) , that is, *Separation with the H-type sending R and the L-type sending C*. If the server adopts this separating strategy, then both R and C are on the equilibrium path, and by Bayes' rule, the customers' beliefs upon observing the signals R and C are updated as $\delta^R = 1$ and $\delta^C = 0$, respectively. Note that the low-quality server will deviate to R if $\lambda^O(1) > \lambda^U(0)$, and the high-quality server will deviate to C if $\lambda^O(1) < \lambda^U(0)$. Therefore, the separating equilibrium $[(R,C), (n(\delta^R), p(\delta^C)), \delta^R = 1, \delta^C = 0]$ can be sustained only if $\lambda^O(1) = \lambda^U(0)$.

(4) (C,R) , that is, *Separation with the H-type sending C and the L-type sending R*. We can similarly argue that the separating equilibrium $[(C,R), (n(\delta^R), p(\delta^C)), \delta^R = 0, \delta^C = 1]$ can be sustained only if $\lambda^O(0) = \lambda^U(1)$.

We denote the unique crossing point of $\lambda^O(\delta)$ and $\lambda^U(\delta)$ as $\hat{\lambda}$. Then, the equilibrium outcomes can be summarized in the following proposition.

Proposition 1. *Consider that all of the customers are uninformed of service quality. Then, if $\lambda < \hat{\lambda}$, the unique pure-strategy perfect sequential equilibrium is pooling on C (i.e., (C,C)); otherwise, it is pooling on R (i.e., (R,R)), except at those potential arrival rates under which $\lambda^O(1) = \lambda^U(0)$ and $\lambda^O(0) = \lambda^U(1)$.*

Proposition 1 indicates that when all of the customers are uninformed, the pure-strategy perfect sequential equilibria must be pooling, except at some specific values of market size (i.e., the potential arrival rate), and thus the equilibrium outcome of the signaling game cannot convey any quality information to the customers. To explain this conclusion, let us first recall the conclusions of both Hassin (1986) and Chen and Frank (2004), who considered a fixed level of service quality and showed that there exists a threshold on the potential arrival rate, below which the server prefers to conceal the queue and above which the server prefers to reveal the queue. This conclusion is based on the assumption that the customers know the server's service quality, and the server's optimal queue disclosure strategy is driven mainly by

the underlying *delay announcement* effect. Also, note that the threshold is quality-level dependent, and thus we shall have two thresholds corresponding to the high- and low-quality servers, respectively. We now consider our setting in which the customers do not know the server's true quality. Then, when the market size is below the minimum (above the maximum) of the two quality-dependent thresholds, both types of servers prefer to conceal (reveal) the queue. When the market size falls between these thresholds, if the server considers only the delay announcement effect, the queue disclosure strategies of these two types of servers will be different: one type of server prefers to conceal the queue, whereas the other type prefers to reveal it. Proposition 1 shows that in this situation, the low-quality server will mimic the high-quality server. Intuitively, mimicking helps the low-quality server to hide its true quality and thus attract more customers to the system.

4.3. Effect of Queue Disclosure as a Signal

We refer to our setting in which the queue disclosure actions have a signaling effect as the *signaling case* and the setting without this effect as the *nonsignaling case*. Under the nonsignaling case, it can be easily verified that the server prefers to conceal the queue if $\lambda^U(\delta) > \lambda^O(\delta)$ and to reveal it otherwise. Accordingly, under the pooling perfect sequential equilibrium, our signaling case performs the same as the nonsignaling case. Thus, by Proposition 1, we have the following result.

Corollary 1. *When all of the customers are uninformed of service quality, the effective arrival rate under the pooling perfect sequential equilibrium is the same as the maximum effective arrival rate under the nonsignaling case.*

Because the pooling strategy is the prevalent equilibrium outcome, using the queue disclosure action as a signal of service quality has no effect on the server's effective arrival rate when all of the customers are uninformed. Does this conclusion still hold if some customers are informed? The existence of informed customers surely affects the server's decision. According to Debo et al. (2012), the queue length can convey some quality information to the uninformed customers when there are both informed and uninformed customers. In this case, it can be more difficult for the low-quality server to mimic the queue disclosure behavior of the high-quality server. Consequently, a separating equilibrium may exist. We investigate this setting with both informed and uninformed customers in the following section.

5. Signaling Game with Heterogeneous Customers

Here, we consider the setting in which some customers are informed of service quality. The informed customers can be either *positively informed* if the server is of high quality or *negatively informed* if the server is of low quality.

We use the variable q ($0 < q < 1$) to represent the fraction of informed customers in the market. The signaling game becomes much more complicated when the customers are heterogeneous. In Section 5.1, we investigate the equilibrium queueing strategies of both informed and uninformed customers in the cases of unobservable and observable queues. In Section 5.2, we study the sequential equilibrium of the whole signaling game. We then examine the signaling effect on the effective arrival rate and customers' total utility in Section 5.3.

5.1. Customers' Equilibrium Queueing Strategies and Effective Arrival Rates

We now analyze customers' equilibrium queueing strategies and derive the corresponding effective arrival rate given the visibility of the queue.

5.1.1. Concealed Queue. First, consider the case in which the server conceals the queue; that is, the queue is unobservable. Assume that upon observing the server's queue concealment behavior, uninformed customers hold a belief that the server's service quality is high with probability δ^C ($0 < \delta^C < 1$)². The uninformed customers then join the system with probability $p_{un}(\delta^C)$. The informed customers who know the server's service quality join the system with probability $p_H(\delta^C)$ ($p_L(\delta^C)$) when the server's service quality is high (low). In such a static game with incomplete information, we denote the customer's queueing strategy by the triplet (p_L, p_{un}, p_H) and the equilibrium strategy profile by (p_L^U, p_{un}^U, p_H^U) , with δ^C omitted in the expression hereafter for notational convenience.

Given the customers' queueing strategy (p_L, p_{un}, p_H) , the expected utility of a positively informed customer is $u_H(p_{un}, p_H) := V_H - \frac{\theta}{\mu - \lambda(qp_H + (1-q)p_{un})}$, of a negatively informed customer is $u_L(p_L, p_{un}) := V_L - \frac{\theta}{\mu - \lambda(qp_L + (1-q)p_{un})}$, and of an uninformed customer is $u_{un}(p_L, p_{un}, p_H) := \delta^C [V_H - \frac{\theta}{\mu - \lambda(qp_H + (1-q)p_{un})}] + (1 - \delta^C) [V_L - \frac{\theta}{\mu - \lambda(qp_L + (1-q)p_{un})}]$. The equilibrium queueing behaviors of all types of customers are determined entirely by the above three utilities. The following proposition gives the customers' equilibrium queueing strategies (p_L^U, p_{un}^U, p_H^U) under various cases. When any one element in the triplet (p_L^U, p_{un}^U, p_H^U) equals 0 or 1, we simply write it as 0 or 1. For example, the triplet $(0, p_{un}^U, 1)$ represents the case where $p_L^U = 0$, $p_{un}^U = 1$, and $0 \leq p_H^U \leq 1$.

Proposition 2. *When the queue is unobservable, the customers' equilibrium queueing strategies (p_L^U, p_{un}^U, p_H^U)*

that hinge on the magnitude of the potential arrival rate and whether λ_1 is smaller than λ_2 are summarized in Table 1, where $\lambda_1 = \frac{\mu - \theta/V_L}{1-q}$, $\lambda_2 = \mu - \frac{\theta}{V_H}$, $\bar{\lambda}$ is the unique value of $\lambda \in (0, \mu)$ that satisfies the equation $u_{un}(0, 1, 1) = 0$, the value of p_L^U in $(p_L^U, 1, 1)$ is $\frac{\mu - \theta/V_L}{\lambda q} - \frac{1-q}{q}$, p_{un}^U in $(0, p_{un}^U, 1)$ is the unique value of $p_{un} \in (0, \min\{\frac{\mu - q\lambda}{(1-q)\lambda}, 1\})$ that satisfies the equation $u_{un}(0, p_{un}, 1) = 0$, and (p_L^U, p_{un}^U, p_H^U) represents a continuum of equilibria with $p_{un}^U \in [\max\{0, \frac{\mu - \theta/V_H}{\lambda(1-q)} - \frac{q}{1-q}\}, \min\{1, \frac{\mu - \theta/V_L}{\lambda(1-q)}\}]$ and the corresponding $p_H^U = \frac{\mu - \theta/V_H}{\lambda q} - \frac{(1-q)p_{un}^U}{q}$ and $p_L^U = \frac{\mu - \theta/V_L}{\lambda q} - \frac{(1-q)p_{un}^U}{q}$.

A close look at the equilibrium outcomes listed in Table 1 reveals that in Case 1 (i.e., when $\lambda_1 < \lambda_2$), as the market size λ increases beyond the point $\frac{\theta(V_H - V_L)}{qV_HV_L}$, the equilibrium outcome evolves from a unique equilibrium to multiple equilibria. Specifically, we can show that when $\lambda < \frac{\theta(V_H - V_L)}{qV_HV_L}$, at least one type of customer has a strictly positive expected utility. Thus, at least one of the three probabilities (p_L^U , p_{un}^U , and p_H^U) is equal to 1, and the others can be uniquely determined by making the corresponding expected utility 0. Subsequently, the final equilibrium triplet is unique. However, when $\lambda > \frac{\theta(V_H - V_L)}{qV_HV_L}$, the expected utilities of all three types of customers are 0 in equilibrium. This leads to three equations, $u_H(p_{un}^U, p_H^U) = 0$, $u_{un}(p_L^U, p_{un}^U, p_H^U) = 0$, and $u_L(p_L^U, p_{un}^U) = 0$, any one of which, however, is redundant given the other two. Consequently, the equilibrium queueing strategy is identified by a system of two nonlinear equations with three variables. Therefore, multiple equilibria exist. Taking into account that the joining probability is between 0 and 1, we can derive that any $p_H^U = \frac{\mu - \theta/V_H}{\lambda q} - \frac{(1-q)p_{un}^U}{q}$ and $p_L^U = \frac{\mu - \theta/V_L}{\lambda q} - \frac{(1-q)p_{un}^U}{q}$ with $p_{un}^U \in [\max\{0, \frac{\mu - \theta/V_H}{\lambda(1-q)} - \frac{q}{1-q}\}, \min\{1, \frac{\mu - \theta/V_L}{\lambda(1-q)}\}]$ is an equilibrium. Note that at the market size $\lambda = \frac{\theta(V_H - V_L)}{qV_HV_L}$, the equilibrium is unique with $p_{un}^U = \frac{qV_HV_L(\mu - \theta/V_L)}{(1-q)\theta(V_H - V_L)}$, $p_H^U = 1$, and $p_L^U = 0$ as $0 < \frac{\mu - \theta/V_H}{\lambda(1-q)} - \frac{q}{1-q} = \frac{\mu - \theta/V_L}{\lambda(1-q)} < 1$. When $\lambda_1 \geq \lambda_2$ (i.e., Case 2), the equilibrium outcome also evolves from a unique equilibrium to multiple equilibria, which can be explained in a similar way.

Furthermore, we observe that in Case 1, the negatively informed customers certainly join in a small market and join with some probability in a large market. However,

Table 1. Equilibrium Queueing Strategies (p_L^U, p_{un}^U, p_H^U) When the Queue Is Unobservable

$\lambda \in$	$(0, \mu - \frac{\theta}{V_L}]$	$(\mu - \frac{\theta}{V_L}, \min(\lambda_1, \lambda_2)]$	$(\min(\lambda_1, \lambda_2), \bar{\lambda}]$	$(\bar{\lambda}, \frac{\theta(V_H - V_L)}{qV_HV_L}]$	$(\frac{\theta(V_H - V_L)}{qV_HV_L}, +\infty)$
Case 1: $\lambda_1 < \lambda_2$	(1, 1, 1)	$(p_L^U, 1, 1)$	(0, 1, 1)	$(0, p_{un}^U, 1)$	(p_L^U, p_{un}^U, p_H^U)
Case 2: $\lambda_1 \geq \lambda_2$	(1, 1, 1)	$(p_L^U, 1, 1)$		(p_L^U, p_{un}^U, p_H^U)	

they never join when the market size falls into an intermediate range (i.e., $\lambda \in \left(\lambda_1, \frac{\theta(V_H - V_L)}{qV_H V_L}\right]$). The underlying reason is as follows: an uninformed customer clearly holds a higher expectation of the level of service quality than a negatively informed customer does. When the market size is moderate, the uninformed customers join the queue with positive probability. This, however, leads to a negative expected utility for the negatively informed customers (i.e., $u_L(0, p_{um}^U) < 0$) and hence, prevents them from joining.

Based on Proposition 2, we can further derive the effective arrival rates under any given market size λ for both high- and low-quality servers. Denote $\lambda_H^U(\delta^C)$ and $\lambda_L^U(\delta^C)$ as the respective effective arrival rates of the high- and low-quality servers when the uninformed customers hold the belief that the server is of high quality with probability δ^C . Then, we get the following results.

Proposition 3. *When the queue is unobservable, $\lambda_H^U(\delta^C)$ and $\lambda_L^U(\delta^C)$, the effective arrival rates of the respective high- and low-quality servers, are as listed in Table 2, where $x(\lambda)$ is the unique value of $x \in (q\lambda, \mu)$ that satisfies the equation $\delta^C V_H + (1 - \delta^C)V_L = \delta^C \frac{\theta}{\mu - x} + (1 - \delta^C) \frac{\theta}{\mu - (x - q\lambda)}$. In Case 1, where $\lambda_1 < \lambda_2$, $\lambda_H^U(\delta^C)$ is nondecreasing with the potential arrival rate λ , and $\lambda_L^U(\delta^C)$ is decreasing with λ when $\lambda \in \left(\bar{\lambda}, \frac{\theta(V_H - V_L)}{qV_H V_L}\right]$ and nondecreasing otherwise. In contrast, in Case 2, where $\lambda_1 \geq \lambda_2$, both $\lambda_H^U(\delta^C)$ and $\lambda_L^U(\delta^C)$ are nondecreasing with λ .*

Although this queueing game may have multiple equilibria when the potential arrival rate falls into certain ranges, Proposition 3 indicates that the effective arrival rates for both types of servers are in fact unique.³ The main reason is that multiple equilibria arise only when all types of customers obtain an expected utility of 0 (see Proposition 2 and its proof). Under such a scenario, different equilibria only affect the composition of the effective arrival rate, that is, the proportion of the joining customers who are informed or uninformed.

Proposition 3 also implies that under Case 1, when the market size falls into the range $\lambda \in \left(\bar{\lambda}, \frac{\theta(V_H - V_L)}{qV_H V_L}\right]$ and the server is of low quality, increasing the market size reduces the effective arrival rate (see Figure 2). This counterintuitive result can be explained as follows.

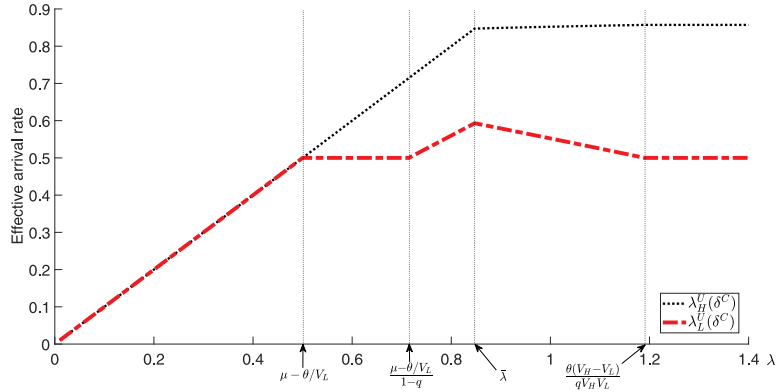
Recall from Proposition 2 that in this situation, the negatively informed customers never join the system. In contrast, uninformed customers do not know the service quality, and they believe that a larger potential arrival rate will bring more positively informed customers to the queueing system with a positive probability. Hence, as the market size increases, to avoid the crowd, uninformed customers will reduce their joining probability p_{um}^U to ensure their expected utility $u_{um}(0, p_{um}^U, 1) = 0$. Consequently, for the low-quality server, as the market size increases within a certain range, the effective arrival rate actually decreases. We call this phenomenon the *low-quality server's market trap*, which, as illustrated in the later analysis, can affect the equilibrium outcome of our signaling game.

5.1.2. Revealed Queue. When the server reveals the queue, all of the customers, both informed and uninformed, inspect the queue length upon arrival and then decide whether to join. Upon observing the server's queue revelation behavior, the uninformed customers hold a belief that the server is of high quality with probability δ^R ($0 < \delta^R < 1$)⁴. For an informed customer, the equilibrium strategy can be easily derived: when the server is of high (low) quality, they join the queue unless it is longer than a threshold $n(1) := \lfloor V_H \mu / \theta \rfloor - 1$ ($n(0) := \lfloor V_L \mu / \theta \rfloor - 1$). In other words, a positively informed customer joins the queue with probability $p_H^O(i) = 1$ at queue length i when $i = 0, 1, \dots, n(1)$ and with probability $p_H^O(i) = 0$ otherwise; a negatively informed customer joins the queue with probability $p_L^O(i) = 1$ at queue length i when $i = 0, 1, \dots, n(0)$ and with probability $p_L^O(i) = 0$ otherwise. The uninformed customers with the belief $\delta^R \in (0, 1)$ can infer the quality information from the queue length (see Debo et al. 2012). Clearly, if the queue is no longer than $n(0)$, an uninformed customer joins the system; that is, the customer's joining probability $p_{um}^O(i)$ is 1 for $i = 0, 1, \dots, n(0)$. Likewise, if the queue length is greater than $n(1)$, the customer balks; that is, the joining probability $p_{um}^O(i)$ is 0 for $i = n(1) + 1, \dots$. We now need to derive the uninformed customers' joining probability at queue length i for $i = n(0) + 1, \dots, n(1)$. We assume that the joining decision of an uninformed customer is made only based on the queue length at arrival (see Debo et al. 2012). In this dynamic game with incomplete information, the customers' equilibrium queueing

Table 2. Effective Arrival Rates $\lambda_H^U(\delta^C)$ and $\lambda_L^U(\delta^C)$ Under a Concealed Queue

$\lambda \in$		$\left(0, \mu - \frac{\theta}{V_L}\right]$	$\left(\mu - \frac{\theta}{V_L}, \min(\lambda_1, \lambda_2)\right]$	$\left(\min(\lambda_1, \lambda_2), \bar{\lambda}\right]$	$\left(\bar{\lambda}, \frac{\theta(V_H - V_L)}{qV_H V_L}\right]$	$\left(\frac{\theta(V_H - V_L)}{qV_H V_L}, +\infty\right)$
Case 1: $\lambda_1 < \lambda_2$	$\lambda_H^U(\delta^C)$	λ			$x(\lambda)$	$\mu - \frac{\theta}{V_H}$
	$\lambda_L^U(\delta^C)$	λ	$\mu - \frac{\theta}{V_L}$	$(1 - q)\lambda$	$x(\lambda) - q\lambda$	$\mu - \frac{\theta}{V_L}$
Case 2: $\lambda_1 \geq \lambda_2$	$\lambda_H^U(\delta^C)$	λ			$\mu - \frac{\theta}{V_H}$	
	$\lambda_L^U(\delta^C)$	λ	$\mu - \frac{\theta}{V_L}$			

Figure 2. (Color online) Effective Arrival Rates of the High- and Low-Quality Servers Under a Concealed Queue: $V_H = 7$, $V_L = 2$, $\mu = 1$, $\theta = 1$, $\delta^C = 0.5$, and $q = 0.3$ ($\lambda_1 < \lambda_2$)



strategy profile can be denoted by a set of the triplet $\{(p_L^O(i), p_{un}^O(i), p_H^O(i))\}_{i=0}^{+\infty}$, with δ^R omitted in the expression for simplicity.

In our game, the two types of servers have the same service rate. It is just a special case of the “consumer game” in Debo et al. (2012) in which different types of servers can adopt different service rates. According to Debo et al. (2012), the pure equilibrium joining strategy of an uninformed customer is a *hole-avoiding* strategy. Specifically, an uninformed customer behaves as a positively informed customer, except at a queue length denoted by n_{hole} (namely, the hole) under which the customer balks. Therefore, the queue-length joining set is $\{0, \dots, n_{hole} - 1, n_{hole} + 1, \dots, n(1)\}$. The underlying reason behind such a hole-avoiding strategy is as follows. Given that all of the uninformed customers behave in this way, the fact that an uninformed customer observes a queue length longer than n_{hole} upon arrival implies that sometime in the past an informed customer had inspected a queue length of n_{hole} and joined the system. The service quality, hence, must be high, because otherwise the informed customer would have balked. Let $\lambda_{i,H}$ ($\lambda_{i,L}$) be the effective arrival rate at queue length i ($i = 0, 1, \dots, n(1) + 1$) and $\pi_{i,H}$ ($\pi_{i,L}$) be the limiting probability that the number of customers in the system equals i when the server is of high (low) quality. For the sake of brevity and space saving, we relegate the detailed review of the uninformed customers’ hole-avoiding decision process to Online Appendix A.3.

Let $\lambda_H^O(\delta^R)$ and $\lambda_L^O(\delta^R)$ be the respective effective arrival rates of the high- and low-quality servers in equilibrium when the uninformed customers hold the belief that the server is of high quality with probability δ^R . According to the time reversibility of the above ergodic birth-and-death (BD) processes, the effective arrival rate is equal to the effective departure rate. Because the departure rate is equal to μ when the system is nonempty, the effective departure rate is $\mu(1 - \pi_{0,H})$ ($\mu(1 - \pi_{0,L})$) for the high-quality (low-quality) server. The two limiting

probabilities of empty system (i.e., $\pi_{0,H}$ and $\pi_{0,L}$) can be calculated according to the two BD processes. For the high-quality server, the actual arrival rate is λ at queue length i ($i = 0, \dots, n_{hole} - 1, n_{hole} + 1, \dots, n(1)$) and $q\lambda$ at queue length n_{hole} , whereas for the low-quality server, the actual arrival rate is λ at a queue length no longer than $n(0)$ and $(1 - q)\lambda$ at queue length i ($i = n(0) + 1, \dots, n_{hole} - 1$). Then, we can obtain the following:

$$\lambda_H^O(\delta^R) = \mu(1 - \pi_{0,H}) = \mu \left(1 - \frac{1}{\sum_{i=0}^{n_{hole}} \rho^i + q \sum_{i=n_{hole}+1}^{n(1)+1} \rho^i} \right); \text{ and}$$

$$\lambda_L^O(\delta^R) = \mu(1 - \pi_{0,L})$$

$$= \mu \left(1 - \frac{1}{\sum_{i=0}^{n(0)+1} \rho^i + \sum_{i=n(0)+2}^{n_{hole}} (1 - q)^{i-n(0)-1} \rho^i} \right).$$

5.2. Sequential Equilibrium Analysis

The sequential equilibrium of our signaling game with heterogeneous customers still needs to satisfy the two conditions stated in Definition 1 (see Section 3.2). Now, the *sequential rationality* condition requires that the customers’ joining rules are (p_L^U, p_{un}^U, p_H^U) and $\{(p_L^O(i), p_{un}^O(i), p_H^O(i))\}_{i=0}^{+\infty}$ and that the server’s signaling rule maximizes its expected payoff such that $\forall t \in \mathbb{T}, f(R | t) > 0$ ($f(C | t) > 0$) only if $\lambda_t^O(\delta^R) \geq \lambda_t^U(\delta^C)$ ($\lambda_t^U(\delta^C) \geq \lambda_t^O(\delta^R)$). Then, we can express a sequential equilibrium in this setting as

$$[(f(R | H), f(R | L)), \{(p_L^U, p_{un}^U, p_H^U), \{(p_L^O(i), p_{un}^O(i), p_H^O(i))\}_{i=0}^{+\infty}\}, \delta^R, \delta^C].$$

Below, we analytically investigate the pure strategies in our signaling game. We also examine whether and when they can be sustained as an equilibrium outcome. The analyses of the hybrid- and mixed-strategy equilibria are relegated to Online Appendix A.4. Again, we have four pure strategies, which are individually specified below.

(1) (R,R) , that is, Pooling on R . Under this strategy, both types of servers choose to always reveal their queues. Then, R is on the equilibrium path, and by Bayes' rule, the uninformed customers' updated belief after observing R is still $\delta^R = \delta$. Let the effective arrival rates of the high- and low-quality servers be $\lambda_H^O(\delta)$ and $\lambda_L^O(\delta)$, respectively. To check whether both types of servers are willing to stay on R , we need to specify the off-equilibrium-path belief δ^C . As long as the off-equilibrium-path belief δ^C leads to $\lambda_H^U(\delta^C) \leq \lambda_H^O(\delta)$ and $\lambda_L^U(\delta^C) \leq \lambda_L^O(\delta)$, both types of servers have no incentive to deviate to C . Hence, with such off-equilibrium-path beliefs, this pooling strategy can be sustained as a sequential equilibrium outcome.

(2) (C,C) , that is, Pooling on C . Under this strategy, both types of servers choose to always conceal their queues. Then, C is on the equilibrium path, and by Bayes' rule, the uninformed customers' updated belief after observing C is still $\delta^C = \delta$. Similarly, as long as the off-equilibrium-path belief δ^R leads to $\lambda_H^O(\delta^R) \leq \lambda_H^U(\delta)$ and $\lambda_L^O(\delta^R) \leq \lambda_L^U(\delta)$, both types of servers have no incentive to deviate to R . Then, with such off-equilibrium-path beliefs, this pooling sequential equilibrium can be sustained.

The following proposition summarizes the sufficient conditions under which the sequential equilibrium is uniquely a pooling one.

Proposition 4. *When the market consists of both informed and uninformed customers, there are two potential-arrival-rate thresholds, $\hat{\lambda}_C$ and $\hat{\lambda}_R$, that satisfy $\hat{\lambda}_R > \hat{\lambda}_C$,⁵ such that when $\lambda < \hat{\lambda}_C$, (C,C) , pooling on C is the unique sequential equilibrium with the off-equilibrium-path belief $\delta^R \in [0, 1]$, and when $\lambda > \hat{\lambda}_R$, (R,R) , pooling on R is the unique sequential equilibrium with the off-equilibrium-path belief $\delta^C \in [0, 1]$.*

Proposition 4 shows that for our signaling game with heterogeneous customers, concealing (revealing) the queue is still the unique dominant strategy for two types of servers when the market size is sufficiently small (large). Thus, separating sequential equilibria can arise only in a medium-sized market.

(3) (R,C) , that is, Separation with the H -type sending R and the L -type sending C . Under this strategy, the high-quality server always reveals the queue, and the low-quality server always conceals the queue. If the server adopts this separating strategy, then both R and C are on the equilibrium path, and by Bayes' rule, the uninformed customers' beliefs upon observing the signals R and C are updated as $\delta^R = 1$ and $\delta^C = 0$, respectively. We now check whether this separating strategy can be sustained as an equilibrium outcome. Note that if $\lambda_L^O(1) > \lambda_L^U(0)$, the low-quality server will become strictly better off by deviating to R , and if $\lambda_H^U(0) > \lambda_H^O(1)$, the high-quality server will benefit by deviating to C . Accordingly,

only when $\lambda_H^O(1) \geq \lambda_H^U(0)$ and $\lambda_L^U(0) \geq \lambda_L^O(1)$ can this separating sequential equilibrium be sustained.

(4) (C,R) , that is, separation with the H -type sending C and the L -type sending R . Under this strategy, the high-quality server always conceals the queue, and the low-quality server always reveals the queue. If the server adopts this separating strategy, then both R and C are on the equilibrium path, and by Bayes' rule, the uninformed customers' beliefs upon observing the signals C and R are updated as $\delta^C = 1$ and $\delta^R = 0$, respectively. Similarly, this separating sequential equilibrium can be sustained only if $\lambda_H^U(1) \geq \lambda_H^O(0)$ and $\lambda_L^O(0) \geq \lambda_L^U(1)$.

Proposition 4 shows that the queue disclosure behaviors of high- and low-quality servers may differ only in a medium-sized market. Suppose that at a certain market size the high-quality (low-quality) server prefers revealing (concealing) the queue in equilibrium, that is, (R,C) . In such a case, the uninformed customers can infer the true quality of the server from the server's queue disclosure action. If the high-quality (low-quality) server deviates to queue concealment (revelation), the server knows that the uninformed customers must believe that it is of low (high) quality upon observing a concealed (revealed) queue. When such a deviation makes the server worse off, the separating sequential equilibrium (R,C) can be sustained. A similar rationale applies to the separating sequential equilibrium (C,R) .

Recall from the above analysis that the separating sequential equilibrium (R,C) can be sustained only if $\lambda_H^O(1) \geq \lambda_H^U(0)$ and $\lambda_L^U(0) \geq \lambda_L^O(1)$, and the separating sequential equilibrium (C,R) can be sustained only if $\lambda_H^U(1) \geq \lambda_H^O(0)$ and $\lambda_L^O(0) \geq \lambda_L^U(1)$. Let $\Lambda_{O \geq U}^{(R,C)} := \{\lambda \mid \lambda_H^O(1) \geq \lambda_H^U(0)\}$, $\Lambda_{U \geq O}^{(R,C)} := \{\lambda \mid \lambda_L^U(0) \geq \lambda_L^O(1)\}$, $\Lambda_{O \geq U}^{(C,R)} := \{\lambda \mid \lambda_H^O(0) \geq \lambda_H^U(1)\}$ and $\Lambda_{U \geq O}^{(C,R)} := \{\lambda \mid \lambda_L^U(1) \geq \lambda_L^O(0)\}$. Then, we can obtain the following results for the separating sequential equilibria.

Proposition 5. *For a medium-sized market with $\lambda \in (\hat{\lambda}_C, \hat{\lambda}_R)$ (except at several threshold points), at most one separating sequential equilibrium — (R,C) or (C,R) — can be sustained. Specifically, the ranges of market size λ in which (R,C) or (C,R) can be sustained are $\Lambda^{(R,C)} := \Lambda_{O \geq U}^{(R,C)} \cap \Lambda_{U \geq O}^{(R,C)}$ and $\Lambda^{(C,R)} := \Lambda_{O \geq U}^{(C,R)} \cap \Lambda_{U \geq O}^{(C,R)}$, respectively.*

Proposition 5 shows the exact ranges of market size ($\Lambda^{(R,C)}$ and $\Lambda^{(C,R)}$) in which a separating sequential equilibrium can arise. Under a separating equilibrium, the server's queue disclosure behavior—revealing or concealing the queue—signals exactly the server's service quality. It is worth mentioning that we cannot rule out the possibility that the pooling and separating equilibria coexist in such medium-sized markets. Note that $\lambda_t^U(d)$ and $\lambda_t^O(d)$ ($t \in \{H, L\}$, $d \in \{0, 1\}$) can be derived explicitly, and thus $\Lambda^{(R,C)}$ and $\Lambda^{(C,R)}$ can be easily identified. For

example, when $(\mu - \theta/V_L)/(1 - q) \geq \mu - \theta/V_H$, based on Proposition 3 and Lemma B1 (given in online Appendix B), it can be easily verified that $\lambda_H^O(1)$ and $\lambda_H^U(0)$ cross at a unique point of λ denoted by $\hat{\lambda}_{H1}$, $\lambda_L^O(1)$ and $\lambda_L^U(0)$ cross uniquely at $\lambda = \hat{\lambda}_{L1}$, $\lambda_H^O(0)$ and $\lambda_H^U(1)$ cross uniquely at $\lambda = \hat{\lambda}_{H0}$, and $\lambda_L^O(0)$ and $\lambda_L^U(1)$ cross uniquely at $\lambda = \hat{\lambda}_{L0}$. If $\hat{\lambda}_{H1} \leq \hat{\lambda}_{L1}$, then (R,C) is the unique separating sequential equilibrium for $\lambda \in [\hat{\lambda}_{H1}, \hat{\lambda}_{L1}]$, and if $\hat{\lambda}_{H0} \geq \hat{\lambda}_{L0}$, then (C,R) is the unique separating sequential equilibrium for $\lambda \in [\hat{\lambda}_{L0}, \hat{\lambda}_{H0}]$. However, when $(\mu - \theta/V_L)/(1 - q) < \mu - \theta/V_H$, the above crossing points may not be unique, and thus $\Lambda^{(R,C)}$ or $\Lambda^{(C,R)}$ may be composed of several disconnected ranges of market size.

Propositions 4 and 5 provide some insights on customers' joining strategy. Note that uninformed customers in our setting have three sources of information to infer the service quality: (1) the prior belief, (2) the signal of the queue disclosure action, and (3) the queue length if it is observable. Proposition 4 shows that in small- and large-sized markets, source 2 does not provide any useful information, and thus uninformed customers rely on sources 1 and 3 to make their joining or balking decision. When the queue is observable, the uninformed customers' equilibrium queueing strategy is a hole-avoiding joining strategy, as demonstrated in Debo et al. (2012). In the case of unobservable queue, uninformed customers' equilibrium queueing strategy is demonstrated in Proposition 2. By contrast, in a medium-sized market where a separating equilibrium exists, source 2 provides the full information about the server's quality type, and thus source 3 becomes redundant. In this situation, an uninformed customer behaves exactly the same as an informed one. These results enrich the conclusion of Debo et al. (2012), who only consider information sources 1 and 3 for uninformed customers in an observable queue.

Recall that when all of the customers are uninformed, the effective arrival rates of the high- and low-quality servers are always the same under any sequential equilibrium (see Section 4.2). However, this result does not hold when the customers are heterogeneous in terms of their possession of information about service quality, as shown in the following corollary.

Corollary 2. *When the market consists of both informed and uninformed customers, the effective arrival rate of the high-quality server is (weakly) larger than that of the low-quality server under any (pure-, mixed-, or hybrid-strategy) sequential equilibrium.*

Clearly, a positively informed customer is always more likely to join a queue than a negatively informed one. An uninformed customer, however, cannot make an exact inference about the server's service quality, and thus the uninformed customer's joining decision is the

same for all types of servers. A combination of the above observations leads to the result in Corollary 2.

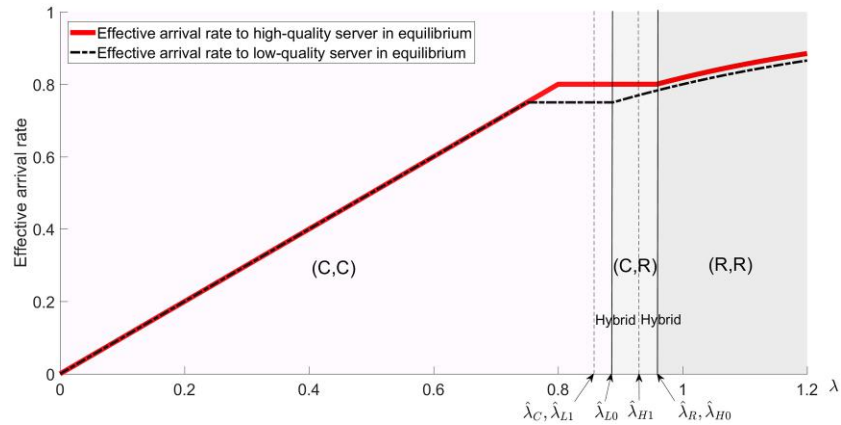
Below, we provide a simple example to illustrate the pure-strategy sequential equilibria. We also numerically examine the hybrid- and mixed-strategy equilibria. For the equilibrium queueing strategy of the uninformed customers in an observable queue, we give priority to the pure strategy with the smallest hole value.

Example 1. Consider the parameter values to be $V_H = 2.5$, $V_L = 2$, $\mu = 1$, $\theta = 0.5$, $\delta = 0.5$, and $q = 0.5$. Under this setting, we have $(\mu - \theta/V_L)/(1 - q) > \mu - \theta/V_H$. Thus, the above-mentioned crossing points are unique, and their values are $\hat{\lambda}_C = \hat{\lambda}_{L1} (= 0.8580) < \hat{\lambda}_{L0} (= 0.8882) < \hat{\lambda}_{H1} (= 0.9265) < \hat{\lambda}_R = \hat{\lambda}_{H0} (= 0.9579)$.

By Proposition 4, we know that the unique sequential equilibrium is (C,C) —pooling on C —with the off-equilibrium-path belief $\delta^R \in [0, 1]$ when the potential arrival rate satisfies $\lambda < \hat{\lambda}_C$ and is (R,R) —pooling on R —with the off-equilibrium-path belief $\delta^C \in [0, 1]$ when $\lambda > \hat{\lambda}_R$. When $\lambda \in [\hat{\lambda}_{L0}, \hat{\lambda}_{H0}]$, according to Proposition 5, the separating equilibrium (C,R) can be sustained. Here, no pooling sequential equilibrium can be sustained (except at the two boundary points). The other separating equilibrium (R,C) does not exist in this example. For the market size $\lambda \in [\hat{\lambda}_C, \hat{\lambda}_{L0})$, we can show that $\lambda_H^O(\delta^R) < \lambda_H^U(\delta)$ always holds for any belief $\delta^R \in [0, 1]$. Therefore, as long as the belief δ^R satisfies $\lambda_L^O(\delta^R) \leq \lambda_L^U(\delta)$, which can be shown to require $0 \leq \delta^R < 1$ in this example, the pooling equilibrium (C,C) can be sustained as a pure-strategy sequential equilibrium outcome. Also, note that here the set $\mathbb{T}' \cup \mathbb{T}''$ is empty. Hence, the credible updating rule does not put any restriction on δ^R . Accordingly, (C,C) is a perfect sequential equilibrium for $\lambda \in [\hat{\lambda}_C, \hat{\lambda}_{L0})$ with the off-equilibrium-path belief $\delta^R \in [0, 1)$. Figure 3 depicts the pure-strategy sequential equilibrium outcome and the corresponding effective arrival rates of both types of servers. It reconfirms Corollary 2 that in equilibrium the effective arrival rate of the high-quality server is always no less than that of the low-quality server.

We now study the hybrid- and mixed-strategy equilibria based on the analysis presented in Online Appendix A.4. First, consider the hybrid strategy $f(R | H) = 1$ and $0 < f(R | L) < 1$. By Bayes' rule, the posterior beliefs of the uninformed customers are $\delta^C = 0$ and $\delta^R = \frac{\delta}{\delta + (1 - \delta)f(R|L)} \in (\delta, 1)$. When $\lambda_L^O(\delta^R) = \lambda_L^U(0)$, we find that the high-quality server strictly prefers to deviate from R to C , improving the server's effective arrival rate from $\lambda_H^O(\delta^R)$ to $\lambda_H^U(0)$. Hence, this hybrid strategy cannot be sustained as an equilibrium outcome. Similarly, the hybrid strategy $0 < f(R | H) < 1$ and $f(R | L) = 0$ cannot be sustained as an equilibrium. Next, consider the hybrid strategy $f(R | H) = 0$ and $0 < f(R | L) < 1$.

Figure 3. (Color online) Sequential Equilibrium Outcome and Corresponding Effective Arrival Rates: $V_H = 2.5$, $V_L = 2$, $\mu = 1$, $\theta = 0.5$, $\delta = 0.5$, and $q = 0.5$



Under this strategy, the posterior beliefs of the uninformed customers are $\delta^R = 0$ and $\delta^C = \frac{\delta}{\delta + (1-\delta)(1-f(C|L))} \in (\delta, 1)$. Only at the unique point $\lambda = \hat{\lambda}_{L0}$ where $\lambda_L^O(0)$ and $\lambda_L^U(\delta^C)$ cross is the low-quality server indifferent between the two choices R and C . It can be verified that the high-quality server has no incentive to deviate at this crossing point, and thus this hybrid strategy can be sustained as an equilibrium outcome at $\lambda = \hat{\lambda}_{L0}$. Similarly, we can show that the hybrid strategy $0 < f(R|H) < 1$ and $f(R|L) = 1$ can be sustained as an equilibrium outcome only at $\lambda = \hat{\lambda}_{H0}$. Last, consider the mixed strategy $0 < f(R|H) < 1$ and $0 < f(R|L) < 1$, which can be sustained as an equilibrium outcome only when $\lambda_t^O(\delta^R) = \lambda_t^U(\delta^C)$ ($t = H, L$). It can be verified that the mixed strategy can never be sustained as a sequential equilibrium outcome in this example.

Next, we conduct the extensive numerical experiments to examine how the changes in the system parameters affect the existence of different kinds of sequential equilibria. Specifically, we vary the following three representative system parameters, the market size λ , the service rate μ , and the monetary reward from the high-quality server V_H , while fixing the other parameter values as $V_L = 2$, $\theta = 0.5$, $\delta = 0.5$, and $q = 0.5$. Note that a larger V_H implies a larger quality gap between the high- and low-quality servers, and the ratio $\rho = \lambda/\mu$ reflects the relative market size in comparison with the server’s capacity. Figure 4 depicts the equilibrium outcomes under various combinations of V_H , λ , and μ . It shows that the pooling equilibrium (C,C) ((R,R)) remains as the unique sequential equilibrium when the market size is sufficiently small (large), and the separating equilibrium (C,R) can arise only in a relatively medium-sized market. Figure 4(a) further shows that the range of market size in which the separating sequential equilibrium can be sustained roughly expands with an increase in V_H . This implies that as the quality gap of two types of servers becomes larger, the

separating equilibrium is more likely to be sustained. Figure 4(b) depicts the equilibrium outcomes when both λ and μ change. From Figure 4(b), we can observe that the region in which a separating equilibrium exists looks roughly diagonal, suggesting that the relative market size (reflected by the ratio ρ) greatly impacts the existence of a separating equilibrium. In Figure 4, (a) and (b), the hybrid equilibrium exists mainly on some boundary curves. This suggests that the region in which a hybrid equilibrium exists is rather limited. The underlying reason is that a hybrid equilibrium can be sustained only at those market-size crossing points where one of two types of servers is indifferent between revealing and concealing the queue.

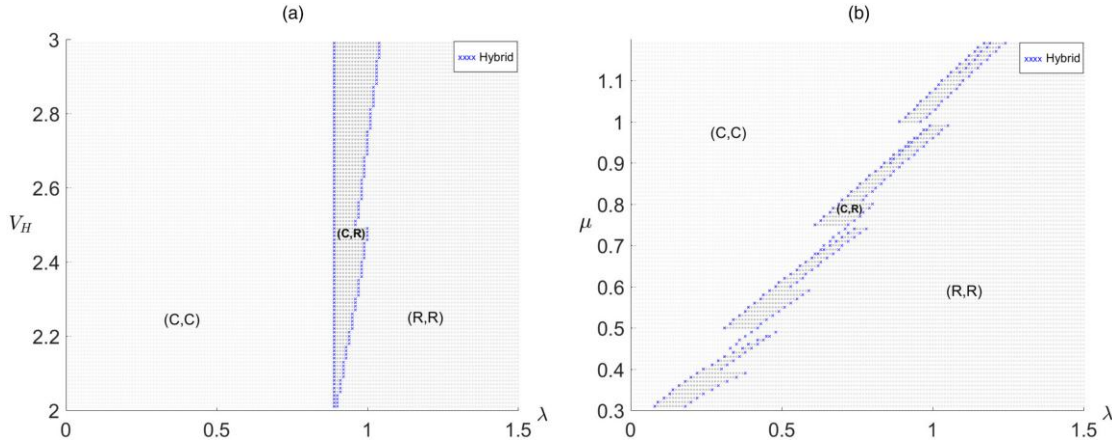
5.3. Effect of Using Queue Disclosure as a Signal

We now examine the effect of using a queue disclosure action as a signal. Here, we focus mainly on the pure-strategy sequential equilibrium because the existence ranges of the hybrid- and mixed-strategy equilibria are very limited (see, e.g., Example 1). In the nonsignaling case, the uninformed customers make their joining decisions based on the expected quality level under their prior belief when the queue is concealed and adopt a “hole-avoiding” strategy that uses both the prior belief and queue length information when the queue is revealed. Anticipating these joining behaviors, the server then makes its queue disclosure decision. Specifically, the t -type ($t = H, L$) server conceals the queue if $\lambda_t^U(\delta) \geq \lambda_t^O(\delta)$ and reveals it if $\lambda_t^U(\delta) \leq \lambda_t^O(\delta)$. Comparing the equilibrium outcome under our signaling case to the outcome under the nonsignaling case, we obtain the following results.

Proposition 6. *When the market is composed of both informed and uninformed customers,*

- (i) *the equilibrium effective arrival rates of both types of servers in the signaling case equal the corresponding maximum ones in the nonsignaling case if the potential arrival rate λ is either smaller than $\hat{\lambda}_C$ or larger than $\hat{\lambda}_R$; and*

Figure 4. (Color online) Impact of System Parameters on the Existence of Different Kinds of Sequential Equilibria: (a) $V_L = 2$, $\mu = 1$, $\theta = 0.5$, $\delta = 0.5$, and $q = 0.5$; and (b) $V_H = 2.5$, $V_L = 2$, $\theta = 0.5$, $\delta = 0.5$, and $q = 0.5$



(ii) for $\lambda \in [\hat{\lambda}_C, \hat{\lambda}_R]$, when a separating sequential equilibrium can be sustained, the maximum effective arrival rate of the high-quality (low-quality) server considering all of the pure-strategy perfect sequential equilibria is no less (no greater) than the maximum one in the nonsignaling case.

In our setting, queue disclosure action has the following two possible effects: the delay announcement effect that can convey the waiting time information to incoming customers and the quality signaling effect that allows uninformed customers to infer the service quality. The delay announcement effect for one type of server is not affected by its counterpart's strategy but is impacted by the magnitude of market size. From Hassin (1986) and Chen and Frank (2004), we know that when the market size is equal to a threshold, revealing and concealing the queue yield the same effective arrival rate; that is, the delay announcement effect is minimal in a medium-sized market. The quality signaling effect, however, is determined by the joint disclosure actions of two types of servers. Propositions 4, 5, and 6 reveal that in both small- and large-sized markets, as the delay announcement effect is strong, it plays a major role, and both types of servers adopt the same queue disclosure action. However, in a medium-sized market, because the delay announcement effect is weak, the quality signaling effect plays a major role instead, allowing the separating equilibrium to arise. We also find that under a separating equilibrium, the low-quality server's queue disclosure action is exactly the same as the one determined by the delay announcement effect. In contrast, the high-quality server's queue disclosure action may not be the same as the one determined by the delay announcement effect, but with the quality signaling effect, the high-quality server still benefits from distinguishing itself from the low-quality server.

Next, we investigate the impact of separating equilibria on the customers' total utility. In a revealed queue where

the belief of the uninformed customers is δ^R , the customers' total utility from a type t server can be derived as

$$u_t^O(\delta^R) = \sum_{i=0}^{n(1)} \lambda_{i,t} \pi_{i,t} \left(V_t - \frac{(i+1)\theta}{\mu} \right), t = H, L. \quad (4)$$

Similarly, in a concealed queue where the belief of the uninformed customers is δ^C , the customers' total utility from a type t server can be written as

$$u_t^U(\delta^C) = \lambda_t^U(\delta^C) \left(V_t - \frac{\theta}{\mu - \lambda_t^U(\delta^C)} \right), t = H, L. \quad (5)$$

Because multiple pure-strategy sequential equilibria may be sustained at the same time in our signaling game, when we derive the customers' total utility from the type t server, we consider by default the pure-strategy perfect sequential equilibrium that brings the maximum effective arrival rate to that type of server (to be consistent with the second statement of Proposition 6). We then have the following result.

Proposition 7. When a separating sequential equilibrium exists, the customers' total utility from a low-quality server is (weakly) larger in the signaling case than in the nonsignaling case.

Proposition 7 indicates that signaling via queue disclosure can make customers better off in a low-quality server system when there is a separating equilibrium. Note that an uninformed customer's expectation of service quality is higher than the low-quality server's quality level. This implies that in the nonsignaling case, the effective arrival rate of the low-quality server is larger when customers are heterogeneous in their possession of quality information than when they are all negatively informed. Such overcrowding results in a negative utility for some of the uninformed customers. In contrast, in the signaling case, under a separating sequential equilibrium, all of the customers become negatively informed

when the server is of low quality. Consequently, the originally uninformed customers become less likely to join the queue, thereby improving their total utility. However, when the server is of high quality, signaling via queue disclosure does not necessarily benefit customers. As all of the customers become positively informed under a separating sequential equilibrium, the joining probability of the uninformed customers increases compared with the nonsignaling case. If such an increase is too large, the system may become overly crowded, which hurts the customer.

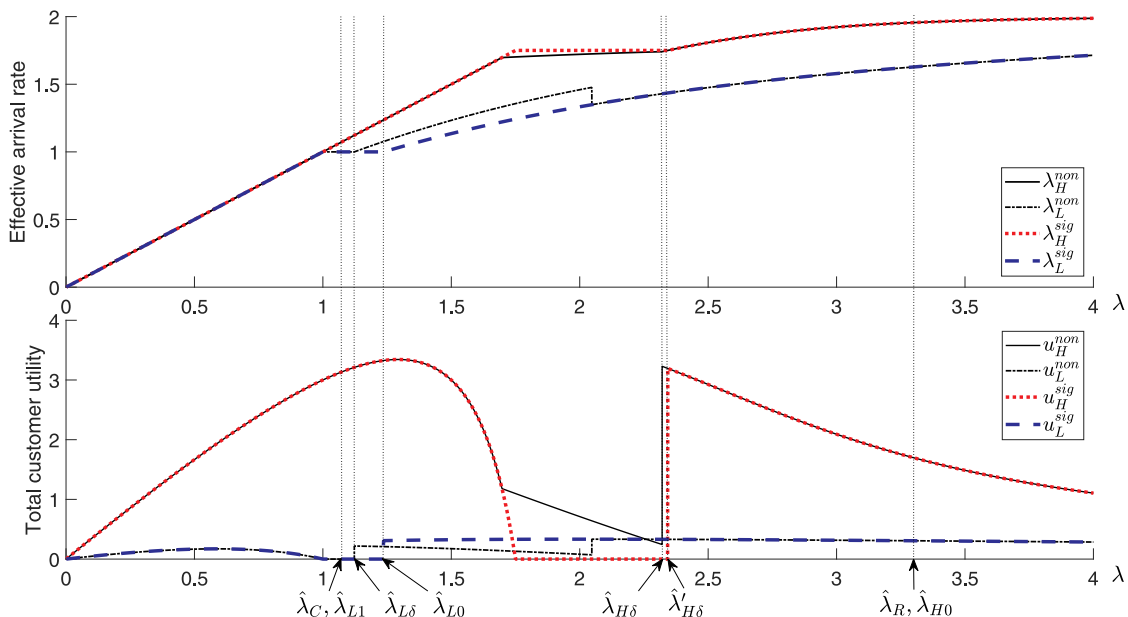
The following numerical example illustrates the above discussions on the effects on effective arrival rates and customers' total utility.

Example 2. Consider the parameter values to be $V_H = 4$, $V_L = 1$, $\mu = 2$, $\theta = 1$, $\delta = 0.25$, and $q = 0.3$. The values of the critical points in Figure 5 are $\hat{\lambda}_C = \hat{\lambda}_{L1} (= 1.0749) < \hat{\lambda}_{L\delta} (= 1.1224) < \hat{\lambda}_{L0} (= 1.2361) < \hat{\lambda}_{H\delta} (= 2.3208) < \hat{\lambda}'_{H\delta} (= 2.3429) < \hat{\lambda}_R = \hat{\lambda}_{H0} (= 3.2997)$.

The upper subfigure in Figure 5 depicts the maximum effective arrival rates of both types of servers in the signaling and nonsignaling cases. In the nonsignaling case, the high-quality server conceals (reveals) the queue when $\lambda \leq \hat{\lambda}_{H\delta}$ ($\lambda > \hat{\lambda}_{H\delta}$), and the low-quality server conceals (reveals) the queue when $\lambda \leq \hat{\lambda}_{L\delta}$ ($\lambda > \hat{\lambda}_{L\delta}$). In the signaling case, by Propositions 4 and 6, we know that only a pooling equilibrium can be sustained as a sequential equilibrium outcome when either $\lambda < \hat{\lambda}_C$ or $\lambda > \hat{\lambda}_R$, and thus the equilibrium effective arrival rates of both types of servers remain unchanged

regardless of whether the queue disclosure action is used as a signaling device or not. When $\hat{\lambda}_{L0} \leq \lambda \leq \hat{\lambda}_{H0}$, the separating sequential equilibrium (C,R) can be sustained. In the subrange $\lambda \in (\hat{\lambda}_{L0}, \hat{\lambda}'_{H\delta})$, (C,R) is the unique pure-strategy perfect sequential equilibrium, and we have $\lambda_H^U(1) > \lambda_H^O(\delta)$ and $\lambda_H^U(1) \geq \lambda_H^U(\delta)$; in the subrange $\lambda \in (\hat{\lambda}'_{H\delta}, \hat{\lambda}_{H0}]$, we have $\lambda_H^O(\delta) > \lambda_H^U(1)$, and the pooling equilibrium (R,R) can also be sustained as a perfect sequential equilibrium with the off-equilibrium-path belief $\delta^C \in [0, 1]$. Therefore, for $\hat{\lambda}_{L0} \leq \lambda \leq \hat{\lambda}_{H0}$, using the queue disclosure action as a signal can make the high-quality server better off and the low-quality server worse off considering all of the pure-strategy perfect sequential equilibria, which confirms the second statement of Proposition 6. For the remaining range $\lambda \in [\hat{\lambda}_C, \hat{\lambda}_{L0})$, only the pooling sequential equilibrium (C,C) can be sustained with the off-equilibrium-path belief δ^R satisfying $\lambda_L^O(\delta^R) \leq \lambda_L^U(\delta)$ (e.g., $\delta^R = 0$), which is also a perfect sequential equilibrium because the credible updating rule puts no restriction on the off-equilibrium-path belief δ^R . For this market size range, the high-quality server conceals the queue in both signaling and nonsignaling cases, and thus the server's optimal effective arrival rates in the two cases are the same. This observation holds for the low-quality server for $\lambda \in [\hat{\lambda}_C, \hat{\lambda}_{L\delta}]$. However, for $\lambda \in (\hat{\lambda}_{L\delta}, \hat{\lambda}_{L0})$, although the signaling effect does not change the belief of the uninformed customers, the effective arrival rate of the low-quality server in the signaling case becomes strictly smaller than the one in the

Figure 5. (Color online) Comparisons of Maximum Effective Arrival Rates of the High-Quality (Low-Quality) Server, λ_H^{non} and λ_H^{sig} (λ_L^{non} and λ_L^{sig}), and the Corresponding Customers' Total Utilities, u_H^{non} and u_H^{sig} (u_L^{non} and u_L^{sig}) in the Nonsignaling and Signaling Cases: $V_H = 4$, $V_L = 1$, $\mu = 2$, $\theta = 1$, $\delta = 0.25$, and $q = 0.3$



nonsignaling case because of the change in the queue disclosure action: the low-quality server reveals the queue in the nonsignaling case but conceals it under the unique pure-strategy perfect sequential equilibrium (C,C) in the signaling case.

Regarding the customers' total utility, the lower subfigure in Figure 5 shows that when the market size λ falls into the range $[\hat{\lambda}_{L0}, \hat{\lambda}_{H0}]$, within which the separating sequential equilibrium (C,R) exists, the customers' total utility from the low-quality server is (weakly) higher in the signaling case than in the nonsignaling case. This confirms the result in Proposition 7. However, when the market size λ falls into the range $(\hat{\lambda}_{L\delta}, \hat{\lambda}_{L0})$, the customers' total utility from the low-quality server becomes strictly lower in the signaling case than in the nonsignaling case because of the low-quality server's different queue disclosure strategies: revealing the queue in the nonsignaling case but concealing it in the signaling case. The lower subfigure indicates that, in this example, the customers' total utility from the high-quality server is (weakly) lower in the signaling case than in the nonsignaling case. In particular, for $\lambda \in [\hat{\lambda}_{L0}, \hat{\lambda}_{H\delta}]$, although the high-quality server conceals the queue in both the signaling and nonsignaling cases, the effective arrival rate is (weakly) larger under the former than under the latter (see the upper subfigure), because the uninformed customers can now fully infer the server's quality in the signaling case. This increase in the system workload hurts the customers' total utility. It is worth mentioning that an increase in the effective arrival rate of the high-quality server does not necessarily harm the consumers.⁶ For $\lambda \in (\hat{\lambda}_{H\delta}, \hat{\lambda}'_{H\delta}]$, the high-quality server conceals the queue in the signaling case but reveals it in the nonsignaling case. In this situation, the customers' total utility is 0 in the former but positive in the latter. Note that the customer's total utility u_H^{sig} jumps upward at $\lambda = \hat{\lambda}'_{H\delta}$. This is because of the change of the server's strategy from concealing to revealing, and here the visibility of the queue benefits the customers.

In the above Example 2, the separating sequential equilibrium that can be sustained is (C,R) . We also present another example where the separating sequential equilibrium that can be sustained is (R,C) and examine its effects. To save space, we relegate it to the Online Appendix A.5.

6. Discussions

We have shown that compared with the nonsignaling case, using a queue disclosure action as a signaling device affects the performance of the system only under a separating sequential equilibrium, which can exist only in a medium-sized market. A separating sequential equilibrium helps the uninformed customers to fully infer the server's quality, leads to a larger effective arrival rate

for the high-quality server, and improves the customers' total utility from the low-quality server. Below, we focus on the separating sequential equilibrium and examine how the composition of two types of customers in the market and service price affect its existence.

6.1. Impact of Customer Type Composition

A close look at the results stated in Section 4.2 and Section 5.2 reveals that the composition of customer types is a critical factor in the existence of a separating equilibrium in our signaling game. A separating equilibrium can be sustained only when the market consists of both informed and uninformed customers. This makes us wonder whether an increase in q , the proportion of informed customers in the market, can make a separating sequential equilibrium more likely to occur.

Recall from Proposition 5 that the ranges of market size λ within which the separating sequential equilibrium can exist critically hinge on the relative magnitudes of eight effective arrival rates denoted by $\lambda_t^U(d)$ and $\lambda_t^O(d)$ ($t \in \{H,L\}, d \in \{0,1\}$). Thus, examining the impact of q on the existence of a separating sequential equilibrium is equivalent to examining its impact on these eight rates. Let $\hat{q} := 1 - \frac{\mu - \theta/V_L}{\mu - \theta/V_H}$, and then we have the following two cases.

Case 1: $q \in [\hat{q}, 1]$. That is, the fraction of uninformed customers $1 - q$ is relatively small. In this situation, it can be verified that the four effective arrival rates in a concealed queue, $\lambda_t^U(d)$, $t \in \{H,L\}, d \in \{0,1\}$, are independent of q regardless of the server's quality (see Online Appendix A.2). We now investigate how an increase in q affects the effective arrival rates in revealed queues, that is, $\lambda_t^O(d)$ ($t \in \{H,L\}, d \in \{0,1\}$).

First, under the separating sequential equilibrium (C,R) , the uninformed customers hold a posterior belief $\delta^R = 0$ ($\delta^C = 1$) when observing a revealed (concealed) queue. Then, all of the customers become negatively informed with a revealed queue, and thus the corresponding effective arrival rate of the low-quality server $\lambda_L^O(0)$ becomes independent of q . That is, the proportion of informed customers does not affect the low-quality server's incentive to stay at R . In contrast, as q increases, more customers become informed, and thus $\lambda_H^O(0)$, the effective arrival rate of the high-quality server if it reveals the queue, becomes larger. This increases the high-quality server's incentive to deviate from C to R , making the equilibrium (C,R) less likely to be sustained. Thus, the range of market size in which the separating sequential equilibrium (C,R) can be sustained, if it exists, becomes smaller as q increases within the range $[\hat{q}, 1]$.

Next, under the separating sequential equilibrium (R,C) , the uninformed customers update their posterior belief as $\delta^R = 1$ ($\delta^C = 0$) when observing a revealed (concealed) queue. Then, all of the customers become positively informed with a revealed queue, and thus the corresponding effective arrival rate of the high-quality

server $\lambda_H^O(1)$ becomes independent of q . This indicates that as q increases in $[\hat{q}, 1]$, the high-quality server's incentive to stay at R remains the same. In contrast, as q increases, more customers become informed and thus $\lambda_L^O(1)$, the effective arrival rate of the low-quality server if the server reveals the queue, becomes smaller. This reduces the low-quality server's incentive to mimic the high-quality server, making the separating equilibrium (R,C) more likely to be sustained. Thus, the range of market size in which the separating sequential equilibrium (R,C) can be sustained, if it exists, becomes larger as q increases in the range $[\hat{q}, 1]$.

Case 2: $q \in (0, \hat{q})$. In this situation, the fraction of uninformed customers in the market is substantially large, and their queueing behavior has a critical impact on the eight effective arrival rates. Now, $\lambda_t^U(d), t \in \{H, L\}$ and $d \in \{0, 1\}$, all depend on q . The effect of increasing q on the existence range of the separating sequential equilibrium becomes uncertain. Take the separating sequential equilibrium (C,R) as an example. For the high-quality server, as q increases, more customers are positively informed and $\lambda_H^O(0)$, the effective arrival rate if the server reveals the queue, becomes (weakly) larger. Hence, the high-quality server has a higher incentive to deviate from C to R , making the equilibrium (C,R) less likely to be sustained. However, for the low-quality server, as q increases, more customers are negatively informed and $\lambda_L^U(1)$, the effective arrival rate if the server conceals the queue, becomes (weakly) smaller. This reduces the low-quality server's incentive to mimic the high-quality server, making the separating equilibrium (C,R) more likely to be sustained. Taken together, this means that increasing q affects the queue disclosure incentives of both the high- and low-quality servers, but in opposite directions. When the former surpasses the latter, the market size range in which the separating sequential equilibrium (C,R) can be sustained, if it exists, becomes smaller,

whereas if the latter dominates, this range becomes larger. Similarly, it can be verified that such an uncertain relationship applies to the separating sequential equilibrium (R,C) .

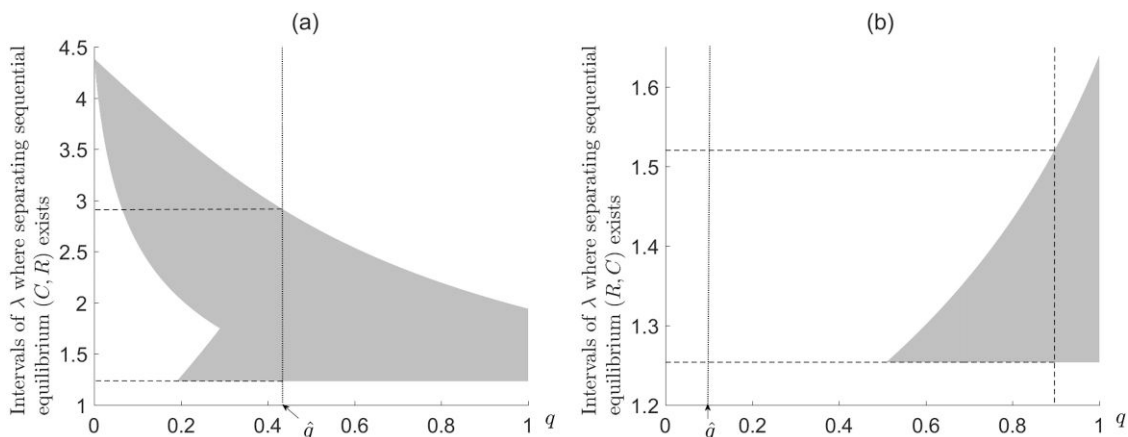
We now use the following example to illustrate the impact of q on the occurrence of separating sequential equilibria.

Example 3. We use two sets of parameter values to illustrate the impact of q on the ranges of market size within which (C,R) and (R,C) can be sustained. First, consider the same set of parameter values used in Example 2, except that we now vary the value of q from 0 to 1. In this case, $\hat{q} = 0.4286$. From Figure 6(a), we can see that the increase of q has a nonmonotonic impact on the existence range of the separating sequential equilibrium (C,R) when $q < \hat{q}$. Specifically, the existence range first increases in q for $q \leq 0.2899$ and then decreases in q for $0.2899 < q < \hat{q}$. Here, the market-size range in which (C,R) can be sustained is discontinuous with two disjoint intervals when q is intermediate-low, leading to a fishtail shape. This is because of the existence of the “low-quality server's market trap” as illustrated in Figure 2. Once q surpasses the threshold \hat{q} , further increasing q makes the equilibrium (C,R) less likely to occur.

Next, consider the parameter values to be $V_H = 1.01, V_L = 0.91, \mu = 2, \theta = 1$, and $\delta = 0.3$ (see also the Online Appendix A.5), and we vary the value of q from 0 to 1. In this scenario, $\hat{q} = 0.1077$. Figure 6(b) shows that the separating sequential equilibrium (R,C) can be sustained only when $q > \hat{q}$. As q increases, the existence range of the separating sequential equilibrium expands.

In practice, high-quality servers often “educate” customers about their service quality through advertising or offering free trials. Here, our results provide a nice money-saving strategy for them: it is unnecessary for a

Figure 6. Impact of q on the Ranges of λ in Which Separating Sequential Equilibria Exist: (a) $V_H = 4, V_L = 1, \mu = 2$, and $\theta = 1$; and (b) $V_H = 1.01, V_L = 0.91, \mu = 2$, and $\theta = 1$



service provider to set a target of educating all customers. Instead, with only a proportion of customers being informed, the remaining uninformed customers can infer the service quality from the service provider's queue disclosure action, which is often costless. Particularly, in the setting of Figure 6(a), the server just needs to set a small q to enable this trick to work over a large range of market size.

6.2. Impact of Service Price

So far, we have assumed that customers receive a monetary reward V_H (V_L) when served by a high-quality (low-quality) server. Note that customers' monetary reward is equal to the service reward minus the service price p . Let the service rewards from the high- and low-quality servers be \mathcal{V}_H and \mathcal{V}_L , respectively. Then, we have $V_H = \mathcal{V}_H - p$ and $V_L = \mathcal{V}_L - p$. We now examine how the service price p affects the existence of a separating sequential equilibrium. Recall that $V_H > V_L > \frac{\theta}{\mu}$ is required to ensure that at least one customer joins the system. Then, the service price p should satisfy the condition $p < \mathcal{V}_L - \frac{\theta}{\mu}$.

Similar to the analysis in Section 6.1, the examination of the impact of p on the existence of separating sequential equilibria can be converted to an examination of its impact on the eight effective arrival rates $\lambda_t^U(d)$ and $\lambda_t^O(d)$, $t \in \{H, L\}$, $d \in \{0, 1\}$. It can be easily verified that these effective arrival rates all decrease in p . Intuitively, the higher the service price p is, the lower the monetary reward is, and thus the less motivated the customers are to join the system. However, the way how each effective arrival rate decreases varies. For $t \in \{H, L\}$ and $d \in \{0, 1\}$, $\lambda_t^U(d)$ decreases in p in a continuous way, whereas $\lambda_t^O(d)$ keeps piecewise constant because of the floor function in $n(0)$ and $n(1)$ and only down jumps (or decreases) at several thresholds of p where $V_t\mu/\theta$ takes integer values. We thus have the following two cases.

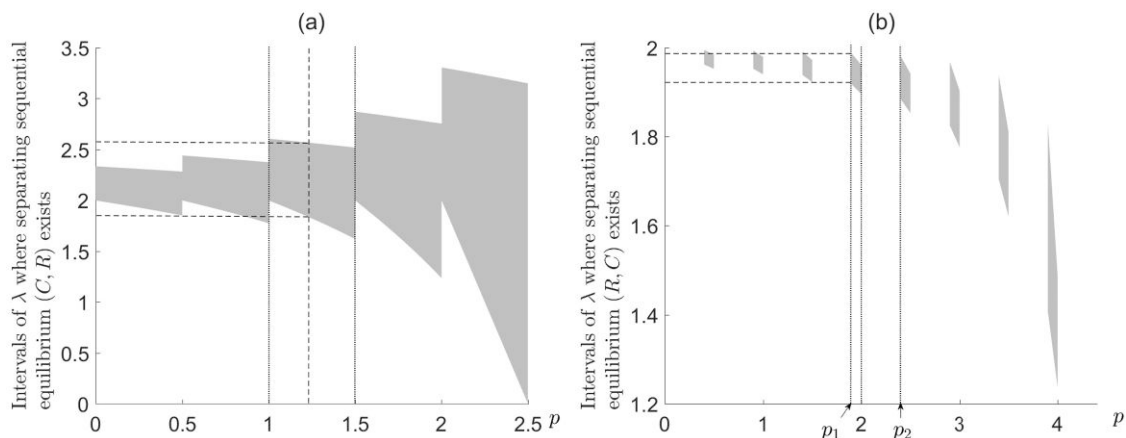
Case 1: Consider each range of the service price p in which both $n(0)$ and $n(1)$ remain unchanged. Then, in each range, $\lambda_t^O(d)$ is a constant. As $\lambda_t^U(d)$ decreases in p , the market size range in which a separating sequential equilibrium exists shifts downward. Intuitively, under the separating sequential equilibrium (C, R) ((R, C)), increasing p attracts fewer customers to join when the queue is concealed. Hence, this makes the low-quality (high-quality) server more likely to reveal the queue in a small-sized market and the high-quality (low-quality) server less likely to conceal the queue in a large-sized market. The combination of the above two effects leads to the downward shift in the range in which the separating equilibrium exists, because p increases in each specific range.

Case 2: Consider the thresholds of p at which $V_H\mu/\theta$ or $V_L\mu/\theta$ takes integer values. In this situation, $\lambda_t^U(d)$ remains almost unchanged, but $\lambda_t^O(d)$ jumps down as p increases across these thresholds. This makes the market size range in which a separating sequential equilibrium exists, if this equilibrium still exists, shift upward at these thresholds.⁷ The underlying reason is similar to the one in Case 1.

We now use the following example to illustrate the impact of service price on the existence of a separating sequential equilibrium.

Example 4. We use two sets of parameter values to illustrate the impact of service price p on the ranges of market size in which a separating sequential equilibrium— (C, R) or (R, C) —exists. First, consider $\mathcal{V}_H = 6$, $\mathcal{V}_L = 3$, $\mu = 2$, $\theta = 1$, and $q = 0.5$. We vary the value of p from 0 to 2.5. In Figure 7(a), the shadowed area represents the ranges of market size in which the separating sequential equilibrium (C, R) exists. It shows that the equilibrium (C, R) always exists when the market size λ falls into a certain interval for all $p \in [0, 2.5)$. Moreover, the range first shifts downward as p increases until it reaches a threshold,

Figure 7. Impact of p on the Ranges of λ in Which Separating Sequential Equilibria Exist: (a) $\mathcal{V}_H = 6$, $\mathcal{V}_L = 3$, $\mu = 2$, $\theta = 1$, and $q = 0.5$; and (b) $\mathcal{V}_H = 5$, $\mathcal{V}_L = 4.9$, $\mu = 2$, $\theta = 1$, and $q = 0.9$



at which point the range abruptly shifts upward. After that, the range again keeps shifting downward and expanding. This pattern keeps repeating itself. This confirms our results stated in Cases 1 and 2.

We can link these observations with the restaurant example to illustrate the managerial implications. In order to attract customers, restaurants in some shopping malls often offer price discounts (or, equivalently, their prices are considered to be low). Our results show that in such a business environment, it might be difficult for a high-quality server to signal the quality via the queue disclosure action. Indeed, when customers are rewarded with an extra gain from price discount, they perhaps care less about the service quality. In contrast, our signaling device might work better for restaurants located in the central business district, which often charge high prices.

Next, consider $\nu_H = 5$, $\nu_L = 4.9$, $\mu = 2$, $\theta = 1$, and $q = 0.9$. We vary the value of p from 0 to 4.4. In Figure 7(b), the disjunct shadowed areas represent the ranges of market size in which the separating sequential equilibrium (R,C) exists. The pattern is similar to the one illustrated in Figure 7(a), except that the equilibrium (R,C) exists only when the service price p falls into several disconnecting intervals. For example, for $p \in (p_1, 2]$ where $p_1 = 1.9$, we have $n(0) = 5$ and $n(1) = 6$, and the shadowed range shifts downward as p increases in this range. Once p is greater than 2, the separating sequential equilibrium (R,C) does not exist until p further increases beyond $p_2 (= 2.4)$. This implies that in some settings, signaling via queue disclosure works only for certain ranges of price.

7. Conclusion

In many service systems, service quality is unknown to some of the incoming customers. The uninformed customers often need to infer the server's service quality level before making their joining or balking decisions. In this study, we consider a signaling game in which the server can signal the service quality through a queue disclosure action, revealing or concealing the queue. We then adopt the sequential equilibrium concept to solve our signaling game and apply the perfect sequential equilibrium as an equilibrium-refinement criterion whenever needed.

We consider the following two scenarios. In the first, all of the customers are uninformed, and in the second, only some of the customers are uninformed. A major finding is that a separating equilibrium exists only when the market size is medium and the system has both informed and uninformed customers. This has multiple implications. First, in a scenario in which all of the customers are uninformed, the pooling perfect sequential equilibrium dominates other equilibria, and thus the queue disclosure action itself cannot convey

any valuable quality information. Second, when customers are heterogeneous in terms of their knowledge of service quality, both types of servers tend to conceal (reveal) their queues when the market size is very small (large), and thus the uninformed customers cannot infer the quality level from the server's queue disclosure action. Third, in a scenario in which a separating equilibrium can be sustained, the server's queue disclosure action fully conveys its service quality information to the uninformed customers. Consequently, the uninformed customers behave exactly the same as the informed customers when making their queueing decisions. The maximum effective arrival rate of the high-quality (low-quality) server is larger (smaller) in our signaling case than in the corresponding case that does not consider the queue disclosure action as a quality signal. Furthermore, this signaling effect of the queue disclosure action can improve the customers' total utility from the low-quality server.

In our study, the server uses the queue disclosure action as a signaling device. In reality, a server can also signal the quality information through other devices such as price; see Debo et al. (2020) for more details. It would be interesting to consider a signaling game in which the server uses price and queue disclosure actions jointly to signal its quality level. We leave this for future research.

Acknowledgments

The authors gratefully thank the department editor (Prof. Morris Cohen), an anonymous associate editor, and two anonymous referees for their very helpful comments and suggestions. All authors contributed equally to the work.

Endnotes

¹ There may exist many posterior beliefs that satisfy the credible updating rule. The *perfect sequential equilibrium* concept, however, does not specify the selection criteria.

² The special cases where δ^C is 0 or 1 are analyzed in Online Appendix A.2.

³ Note that this result holds for the case $0 < \delta^C < 1$ only. When $\delta^C = 0$ or 1, the uniqueness no longer holds. For convenience and to ensure consistency in the sequential equilibrium analysis, we restrict our attention to $\lambda_H^U(0) = \lim_{\delta^C \rightarrow 0^+} \lambda_H^U(\delta^C)$ and $\lambda_L^U(1) = \lim_{\delta^C \rightarrow 1^-} \lambda_L^U(\delta^C)$ by the continuities of $\lambda_H^U(\delta^C)$ and $\lambda_L^U(\delta^C)$ in δ^C ($0 < \delta^C < 1$). The related analysis can be found in Online Appendix A.2.

⁴ The two special cases where δ^R is either 0 or 1 are analyzed in Lemma B1 in Online Appendix B.

⁵ The detailed expressions of $\hat{\lambda}_C$ and $\hat{\lambda}_R$ can be found in the proof of Proposition 4.

⁶ For example, when the parameter values are $V_H = 10$, $V_L = 1$, $\mu = 2$, $\theta = 1$, $\delta = 0.01$, and $q = 0.3$, at the market size $\lambda = 1.5450$, the high-quality server conceals the queue in both the signaling and non-signaling cases. The corresponding effective arrival rates are $\lambda_H^U(1) = 1.5451$ and $\lambda_H^U(\delta) = 1.5370$, and the corresponding customers' total utilities are $u_H^U(1) = 12.0544$ and $u_H^U(\delta) = 12.0503$, respectively. In this situation, the customers' total utility from the high-quality server is higher in the signaling case than in the non-signaling case, that is, $u_H^U(1) > u_H^U(\delta)$.

⁷ It is worth mentioning that the bulge shape of $\lambda_L^U(1)$ (see Figure 2) may narrow down some subranges of λ in which (C, R) can be sustained as a separating sequential equilibrium. However, the overall effect at these thresholds of p is still the upward shift of the range.

References

- Aksin Z, Armony M, Mehrotra V (2007) The modern call-center: A multi-disciplinary perspective on operations management research. *Production Oper. Management* 16(6):665–688.
- Allon G, Bassamboo A, Gurvich I (2011) “We will be right with you”: Managing customer expectations with vague promises and cheap talk. *Oper. Res.* 59(6):1382–1394.
- Armony M, Maglaras C (2004a) Contact centers with a call-back option and real-time delay information. *Oper. Res.* 52(4):527–545.
- Armony M, Maglaras C (2004b) On customer contact centers with a call-back option: Customer decisions, routing rules, and system design. *Oper. Res.* 52(2):271–292.
- Armony M, Shimkin N, Whitt W (2009) The impact of delay announcements in many-server queues with abandonment. *Oper. Res.* 57(1):66–81.
- Burnetas A, Economou A (2007) Equilibrium customer strategies in a single server Markovian queue with setup times. *Queueing Syst.* 56(3–4):213–228.
- Chen H, Frank M (2004) Monopoly pricing when customers queue. *IEE Trans.* 36(6):1–13.
- Debo L, Veeraraghavan S (2014) Equilibrium in queues under unknown service times and service value. *Oper. Res.* 62(1): 38–57.
- Debo L, Parlar C, Rajan U (2012) Signaling quality via queues. *Management Sci.* 58(5):876–891.
- Debo L, Rajan U, Veeraraghavan S (2020) Signaling quality via long lines and uninformative prices. *Manufacturing Service Oper. Management* 22(3):513–527.
- Edelson N, Hildebrand D (1975) Congestion tolls for Poisson queueing processes. *Econometrica* 43(1):81–92.
- Grossman SJ, Perry M (1986) Perfect sequential equilibrium. *J. Econom. Theory* 39(1):97–119.
- Guo P, Hassin R (2011) Strategic behavior and social optimization in Markovian vacation queues. *Oper. Res.* 59(4):986–997.
- Guo P, Zipkin P (2007) Analysis and comparison of queues with different levels of delay information. *Management Sci.* 53(6):962–970.
- Guo P, Haviv M, Luo Z, Wang Y (2022) Optimal queue length information disclosure when service quality is uncertain. *Production Oper. Management* 31(5):1912–1927.
- Hassin R (1986) Consumer information in markets with random product quality: The case of queues and balking. *Econometrica* 54(5):1185–1195.
- Hassin R (2007) Information and uncertainty in a queueing system. *Probab. Engrg. Inform. Sci.* 21(3):361–380.
- Hassin R (2016) *Rational Queueing* (Chapman and Hall/CRC, London).
- Hassin R, Haviv M (1994) Equilibrium strategies and the value of information in a two line queueing system with threshold jockeying. *Stochastic Models* 10(2):415–435.
- Hassin R, Haviv M (2003) *To Queue or not to Queue: Equilibrium Behavior in Queueing Systems* (Kluwer Academic Publishers, London).
- Hassin R, Roet-Green R (2017) The impact of inspection cost on equilibrium, revenue, and social welfare in a single-server queue. *Oper. Res.* 65(3):804–820.
- Hassin R, Roet-Green R (2018) Cascade equilibrium strategies in a two-server queueing system with inspection cost. *Eur. J. Oper. Res.* 267(3):1014–1026.
- Hu M, Li Y, Wang J (2018) Efficient ignorance: Information heterogeneity in a queue. *Management Sci.* 64(6):2650–2671.
- Ibrahim R (2018) Sharing delay information in service systems: A literature survey. *Queueing Syst.* 89(1–2):49–79.
- Ibrahim R, Armony M, Bassamboo A (2017) Does the past predict the future? The case of delay announcements in service systems. *Management Sci.* 63(6):1762–1780.
- Kamenica E, Gentzkow M (2011) Bayesian persuasion. *Amer. Econom. Rev.* 101(6):2590–2615.
- Kremer M, Debo L (2016) Inferring quality from wait time. *Management Sci.* 62(10):3023–3038.
- Kreps DM, Wilson R (1982) Sequential equilibria. *Econometrica* 50(4):863–894.
- Lingenbrink D, Iyer K (2019) Optimal signaling mechanisms in unobservable queues. *Oper. Res.* 67(5):1397–1416.
- Naor P (1969) The regulation of queue size by levying tolls. *Econometrica* 37(1):15–24.
- Veeraraghavan S, Debo L (2009) Joining longer queues: Information externalities in queue choice. *Manuf. Serv. Oper. Manag.* 11(4):543–562.
- Veeraraghavan S, Debo L (2011) Herding in queues with waiting costs: Rationality and regret. *Manufacturing Service Oper. Management* 13(3):329–346.
- Wang J, Hu M (2020) Efficient inaccuracy: User-generated information sharing in a queue. *Management Sci.* 66(10):4648–4666.
- Whitt W (1999) Improving service by informing customers about anticipated delays. *Management Sci.* 45(2):192–207.
- Yu M, Debo L, Kapuscinski R (2016) Strategic waiting for consumer-generated quality information: Dynamic pricing of new experience goods. *Management Sci.* 62(2):410–435.
- Yu Q, Allon G, Bassamboo A (2017) How do delay announcements shape customer behavior? An empirical study. *Management Sci.* 63(1):1–20.
- Yu Q, Allon G, Bassamboo A (2021) The reference effect of delay announcements: A field experiment. *Management Sci.* 67(12): 7417–7437.
- Yu Q, Allon G, Bassamboo A, Iravani S (2018) Managing customer expectations and priorities in service systems. *Management Sci.* 64(8):3942–3970.