Optimal Queue Length Information Disclosure When Service Quality is Uncertain

Pengfei Guo

Department of Management Sciences, City University of Hong Kong, Hong Kong, penguo@cityu.edu.hk

Moshe Haviv

School of Data Science, The Chinese University of Hong Kong, Shenzhen Campus, China, and Department of Statistics and Data Science and the Federmann Center for the Study of Rationality, The Hebrew University of Jerusalem, Israel, moshe.haviv@gmail.com

Zhenwei Luo

Faculty of Business, The Hong Kong Polytechnic University, Hong Kong, zhen-wei.luo@polyu.edu.hk

Yulan Wang

Faculty of Business, The Hong Kong Polytechnic University, Hong Kong, yulan.wang@polyu.edu.hk

---

Abstract. We investigate a server's best queue disclosure strategy in a single-server service system with an uncertain quality level (which is assumed to be binary). We consider this problem from the perspective of a Bayesian persuasion game. The server first commits to a possibly mixed strategy stating that given a realized quality level, whether or not the queue length will be revealed to customers upon their arrival. The service quality level is then realized, and the server's corresponding queue-disclosure action is observed by customers, who then update their beliefs regarding service quality and decide whether or not to join the service system. We then reformulate the server's decision problem as looking for the best Bayes-plausible distribution of posterior beliefs regarding service quality. We demonstrate that the maximal expected effective arrival rate, as a function of the prior belief, can be graphed as the upper envelope of all convex combinations of any two arbitrary points on the two effective arrival rate functions of the revealed and concealed queues. We show that when the market size is sufficiently small (large), the server always conceals (reveals) the queue, regardless of the realized service quality. Numerically, we find that in a medium-sized market, the server's optimal commitment strategy is often hybrid or mixed, that is, randomizing queue concealment and revelation. We also extend our analysis to a situation in which the server aims to maximize social welfare. We show that under certain conditions, it is always beneficial for the welfare-maximizing social planner to randomize queue concealment and revelation, regardless of the market size.

Keywords: Service Quality; Queue Disclosure; Queue Concealment; Queueing Game; Bayesian Persuasion

---

1 Introduction

Whether to provide queue length information to customers, who in turn decide whether or not to join the queue, is a classic research topic. It is well documented in the queueing game literature (e.g., Hassin and Haviv, 2003, p. 51) that there exists a threshold on the arrival rate, below which the server prefers concealing the queue length and above which he prefers revealing it. This queue disclosure strategy is based on a setting with a known service quality; however, in real practice, the quality of provided service may be uncertain. For example, in a restaurant, the food quality may vary due to factors such as the freshness of ingredients and skill of the chef. The quality of online consulting services, such as online healthcare diagnosis platforms and telephone hotlines, heavily relies on the skills and expertise levels of the consultants and agents, which represent a source of uncertainty to customers, particularly at times of rotation in these professionals' schedules. We therefore have the following research question: in service systems with uncertain service quality, how should the server select his queue disclosure strategy?

The server can choose to either reveal or conceal the queue, regardless of the realized service quality. Under such a quality-independent queue-disclosure strategy, customers can choose to join the queue or balk based on their prior beliefs about the service quality. We use the following example to illustrate this strategy.

Example 1. (Quality-Independent Queue-Disclosure Strategy) *A server (he) provides a service with an uncertain quality level. Nature decides whether the service quality is high, with a value of* 2*, or low, with a value of* 1*, according to a Bernoulli trial with respective probabilities of* 0.33 *and* 0.67*. Customers' service times follow an exponential distribution with a rate of* 1.1*. Potential customers arrive according to a Poisson process with a rate of* 0.5*. The waiting cost per unit time is* 1*. Both the server and the customers hold the same prior belief regarding the service quality.*

When the queue length is always revealed to the customers, the customers' belief regarding service quality is 1.33*, the expected one. In this scenario, customers join the observable queue if and only if the queue is empty, and the effective arrival rate can be calculated to be* 0.3438*. Similarly, when the queue is always concealed from customers, the customers' belief regarding service quality is still* 1.33*, the expected one. The effective arrival rate can be derived by setting the customer joining utility to zero, as the potential arrival rate is sufficiently high, yielding an equilibrium arrival rate of* 0.3481*. Hence, the server's optimal quality-independent queue-disclosure strategy is concealing the queue length.*

The quality-independent queue-disclosure strategy, however, is not optimal from the perspective

of arrival-rate maximization for the server. In reality, customers can tolerate a higher level of congestion when expecting a higher level of service quality. Using this information, the server can tailor his queue-disclosure strategy according to the realized service quality to attract more customers to join the queue. Continuing with the above example, consider that the server adopts a *quality-dependent pure queue-disclosure strategy* in which he conceals the queue when service quality is high and reveals it otherwise, and assume that this strategy is announced to customers ex ante. Then, the incoming customer can exactly infer the service quality by the visibility of the queue and do not rely on the expected quality to inform queueing decisions. We can show that this quality-dependent pure queue-disclosure strategy yields an increase in the expected effective arrival rate to 0.3953, a 13.56% improvement over the rate when the server always conceals the queue. Compared with the quality-independent queue-disclosure strategy, this quality-dependent queue-disclosure strategy yields a better match between the realized quality level and the queue-disclosure action, leading to a larger expected arrival rate. We name this outcome the *differentiation effect* of a quality-dependent queue-disclosure strategy.

The above quality-dependent strategy requires quality information to be credibly conveyed to customers via a pre-commitment mechanism. That is, the server must announce his queue disclosure strategy before the service quality is realized; once the quality level is realized, the corresponding queue-disclosure action must be performed without manipulation. Following this idea, we consider a general version of a quality-dependent queue-disclosure strategy. This general strategy can be characterized by two conditional queue-disclosure probabilities, $\pi(\cdot|high)$ and $\pi(\cdot|low)$, which correspond to a high or low level of realized service quality, respectively. The server then commits to this strategy before the service quality is realized. The foregoing quality-dependent pure queue-disclosure strategy can be expressed as $\pi(concealing|high) = 1$ and $\pi(revealing|low) = 1$ and clearly represents a special case of the general strategy. As we show in Section 4.1, the optimal quality-dependent queue-disclosure strategy for Example 1 is a randomization strategy characterized by

$$\pi(revealing|high) = 0,\ \pi(revealing|low) = 0.7537;$$

$$\pi(concealing|high) = 1,\ \pi(concealing|low) = 0.2463.$$

That is, when the realized service quality is low, the server randomizes queue length revelation and concealment, instead of fully disclosing the queue length (as discussed above), with the probability of revealing the queue being 0.7537. At a probability of 0.5050 (0.4950), the queue is revealed (concealed) to the customers. Thus, the posterior probability for the service quality to be high becomes 0 (0.6667), leading to an effective arrival rate of 0.3438 (0.5). Consequently, the effective arrival rate associated with a concealed queue is still 0.5, but the probability of its occurrence increases from

0.33 to 0.4950. The expected effective arrival rate is now 0.4211, a further improvement of 6.53% over the rate achieved using the quality-dependent pure queue-disclosure strategy. This experimental result is quite insightful, as it demonstrates that partial disclosure of quality via strategic randomization can be beneficial to the server. It indicates that in addition to the *differentiation effect*, another benefit is associated with the quality-dependent queue-disclosure strategy; we name this the *persuasion effect*.

In this study, we aim to illustrate the mechanism explaining why a quality-dependent queue-disclosure strategy (particularly the randomization strategy) can yield a larger effective arrival rate for the server. We also aim to provide an approach to find such an optimal strategy. One way to interpret pre-commitment to the randomized queue-disclosure strategy is to consider a "long-run" server who aims to maximize his long-run average profit when facing "short-run" customers (see, e.g., Rayo and Segal, 2010). Customers can then infer the server's randomization strategy from their long-term experiences with and observations of the server's queue-disclosure actions. Note that advances in information technology have made it relatively easy for service providers to change the visibility status of a queue. Using phone-call systems as an example, servers can choose to play music or provide queue-length information to waiting customers. Similarly, servers can also easily control the provision of real-time queue information on online platforms and mobile apps. The quality-dependent queue-disclosure strategy considered in our study can help service providers to persuade more customers to join their system. When well-constructed, such a strategy can also help a welfare-maximizing social planner to better regulate customer arrivals.

Specifically, we consider a stylized single-server service system in which customers arrive according to a Poisson process, and service times are exponentially distributed. However, the service quality provided by the server is random and takes the value of being either high (labeled as h) or low (labeled as l). The realized quality level is available to the server but not the customers. The probability of the service quality being high is common knowledge and hence forms customers' prior belief regarding service quality. Customers are homogeneous: they have the same prior knowledge, receive the same service rewards, and incur the same unit-time delay cost. Before realizing the service quality level, the server announces and commits to his queue disclosure strategy, which is characterized by two conditional probabilities of revealing the queue length at two possible levels of realized service quality. Once the service quality is realized, he then performs the corresponding queue-disclosure action. Based on the visibility of the queue, customers update their beliefs about the service quality according to Bayes' rule and decide to join the queue or balk accordingly to maximize the expected utility.

Directly maximizing the server's expected effective arrival rate does not yield a closed-form

solution. Inspired by Kamenica and Gentzkow (2011), we use a geometric approach to derive the optimal disclosure strategy. We first plot the effective arrival rates of the revealed and concealed queues as two functions of the probability of the service quality being high. We then demonstrate that any convex combination of two points from these two functions can be generated through a properly designed queue disclosure strategy. We further show that any point on the upper envelope of all convex combinations represents the maximal effective arrival rate under the corresponding prior belief. Graphically, we can then determine whether the server benefits from the randomized queue-disclosure strategy by simply checking whether the upper envelope is located strictly above the two effective arrival rate functions.

After deriving the optimal disclosure strategy, we examine the effect of market size (i.e., the potential arrival rate) on the optimal disclosure strategy. We show that when the market size is sufficiently small, the server always conceals the queue length information, regardless of the level of realized service quality. In contrast, when the market size is very large, the server always reveals the queue. These two results are consistent with those identified in studies involving a known level of service quality (e.g., Hassin and Haviv, 2003, p. 51). However, when the market is medium-sized, numerically we find that it is often optimal for the server to adopt a quality-dependent queue-disclosure strategy, as this can help to increase the server's effective arrival rate. Moreover, such a strategy is often hybrid or mixed, that is, queue concealment and revelation actions are randomized.

We further extend our analysis to a setting in which the server acts as a social planner and aims to maximize social welfare. We show that we can still apply our geometric approach to determine the best queue disclosure strategy in this setting. In contrast to the classic finding that revealing the queue length is always socially optimal (e.g., Hassin and Haviv, 2003; Hassin and Roet-Green, 2017), we find that when service quality is uncertain, a randomized queue-disclosure strategy can benefit the social planner.

The remainder of the paper is organized as follows. Section 2 reviews the related literature. The formal model is presented in Section 3. We investigate the optimal queue disclosure strategy in Section 4. In Section 5, we examine a situation in which the server acts as a social planner. Finally, we provide our concluding remarks in Section 6. All of the proofs are relegated to the Appendix.

2 Literature Review

Our work is closely related to studies on quality disclosure. In economics, Grossman (1981) investigates product quality disclosure problems via ex post verifiable disclosures and warranties, showing that the seller voluntarily discloses private information in equilibrium if the disclosure is

costless and information is verifiable. Milgrom (1981) characterizes the favorableness of news and introduces the novel persuasion game, showing that in a sales encounter model, a salesman always reports the most favorable data about his product. In operations management, there are many studies on quality signaling games involving queues. Veeraraghavan and Debo (2009, 2011) consider the quality issue in a two-parallel-observable-queue setting and show that in equilibrium, it might be optimal for customers to join a longer queue. Debo et al. (2012) consider both informed and uninformed customers under an observable queue setting. Since the customers are heterogeneous in terms of possessing information about service quality, the queue length serves as a quality signal to uninformed customers. Some recent works consider other quality signals, such as service or waiting times (Debo and Veeraraghavan, 2014; Kremer and Debo, 2016), information generated by customers (Yu et al., 2016), and price and wait lines (Debo et al., 2020). Guo et al. (2020) consider a situation in which the queue-disclosure behavior can serve as a signaling device, and derive pooling and separating equilibria in this setting. In particular, they find that if the system only has uninformed customers, only the pooling equilibria exist, and thus queue disclosure cannot convey quality information to customers. We note some differences between a quality-signaling game and our queue-disclosure game. A signaling game considers a server with a fixed level of service quality, whereas our setting contains a server with an uncertain level of service quality. These games also differ in terms of timing: in a signaling game, the server has a given quality type, which he signals to customers. However, in our persuasion game, the server commits to a queue disclosure strategy *before* his quality type is realized and must remain committed to this strategy once the quality is realized; a signaling game has no such requirement.

The solution technique we use in this study is closely related to that used in *Bayesian persuasion* games as introduced by Kamenica and Gentzkow (2011), who study how a sender designs a signaling system and commits to it to induce preferred actions from an information receiver. Kamenica and Gentzkow (2011) demonstrate that concavification of the value function identifies whether the sender benefits from persuasion but observe that the structure of the optimal signal can be difficult to derive when the state space is large. Gentzkow and Kamenica (2016) show the optimal signal structure of a particular class of Bayesian persuasion games in which the receiver's optimal action depends only on expectation of the unknown state and the sender's payoff is independent of the state. Lingenbrink and Iyer (2019) first introduce the Bayesian persuasion game into a queueing setting in which the only unknown state of the world is the queue length, and prove that the optimal signaling mechanism is a queue-length-dependent binary threshold signal. Different from Lingenbrink and Iyer (2019), in which the level of service quality is given, we consider a queueing setting with uncertain service quality. This uncertainty leads to both a differentiation effect and

persuasion effect associated with the queue disclosure strategy. The former effect concerns a better match between the realized quality and queue-disclosure action, and the latter effect is related to the partial disclosure of quality information to customers. Our methodology and focus are also different from those of Lingenbrink and Iyer (2019), who use a linear programming model to find the best signal. In contrast, by restricting the signals to queue concealment and revelation, we use a geometric approach to intuitively illustrate the benefits of a quality-dependent queue-disclosure strategy.

Our work is also related to studies on information provision and purchase in queues. Hassin and Haviv (1994) consider a case in which customers arriving at two parallel queues can choose to purchase information about queue length to enable them to join the shorter queue. Hassin (2007) examines a scenario in which service quality and some other system parameters are known to the server but not to the customers, and the server can choose a profit-maximizing price and also can determine whether or not to disclose his private information to the customers. The paper shows that informing customers about the realized system parameters does not necessarily benefit the server or the social planner. Hassin and Roet-Green (2017) study information purchase in a one-server queue setting in which incoming customers can purchase information about the queue length. Later, Hassin and Roet-Green (2018) consider a setting in which customers arriving at parallel servers use the queue length of one server to deduce whether to join the queue or inspect another queue. In this setting, those who have inspected other queues act as informed customers, and the fraction of informed customers is not predetermined but is rather an artifact of the customers' strategy choices.

The research on the effects of delay announcements on queues is also related. Allon et al. (2011) consider a *cheap talk* game between the server and customers in which the server knows the state of the system and sends a related signal, and the customers use this signal to update their belief regarding the expected waiting time. Our work differs from Allon et al. (2011) in two ways: the sender does not commit to a signaling rule under a cheap talk game while our ex-ante commitment approach does. Besides, in Allon et al. (2011), they do not consider quality issues while we do. Yu et al. (2018) further study a cheap talk game in a setting with heterogeneous customers and show that customers' responses to a delay announcement can be used to elicit information on customer type. Other related studies in this stream include Hassin (1986), 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), Yu et al. (2017), Hu et al. (2018) and Yu et al. (2021). We further refer interested readers to two survey books, Hassin and Haviv (2003) and Hassin (2016), and survey papers by Aksin et al. (2007) and Ibrahim (2018) for more studies in this research stream. In a recent study, Li et al. (2020) explore an optimal queue disclosure

strategy with a known level of service quality and demonstrate that it is socially optimal to disclose the queue length only if the queue is either very short or very long. In contrast, we consider the server's optimal queue disclosure strategy when service quality is uncertain, and our queue disclosure strategy is quality-dependent.

Interestingly, our conclusion on the optimal queue disclosure strategy with respect to market size is similar to the findings of Hassin and Roet-Green (2017) and Hu et al. (2018). We all find that to maximize the effective arrival rate, the queue length should be concealed in a small-sized market and revealed in a large-sized market, whereas partial queue disclosure is optimal in a medium-sized market. However, the settings and driving forces behind our conclusions are quite different. In Hassin and Roet-Green (2017), partial information disclosure is achieved by imposing an inspection cost, whereas in Hu et al. (2018), it is achieved by informing only some customers. In contrast, in our setting, partial queue disclosure is scenario-based: according to the realized service quality, incoming customers are either all informed or all uninformed about the queue length according to the server's pre-determined probabilities. In Hassin and Roet-Green (2017) and Hu et al. (2018), customer-based information disclosure helps the server to extract the surplus of customers in certain conditions by exploiting heterogeneity in the customer group. In our work, scenario-based information disclosure has two effects: the differentiation effect achieved by better matching the realized quality level with the queue-disclosure action, and the persuasion effect achieved by manipulating customers' posterior beliefs about the uncertain state (i.e., quality level).

3 Model Description

Consider a single-server queueing system. Potential customers arrive according to a Poisson process at a rate of λ. Their service times follow an exponential distribution with mean $1/\mu$. Let $\rho := \lambda/\mu$. The level of service quality can be high, V_h, with probability δ_0 or low, V_l, with probability $1-\delta_0$. All customers, upon joining the queue, receive the same quality of service and incur a waiting cost of θ per unit of time. We require $V_h > V_l > \frac{\theta}{\mu}$ to ensure that at least one customer joins the system. Customers make their joining-or-balking decisions to maximize their own utility. Nature determines the level of service quality, and the lottery is done once. All of the above information is known to both the server and customers. Before the realization of service quality, the server decides his queue length disclosure strategy by selecting two conditional probabilities, f_h and f_l, that represent the probability that the queue length information is revealed to all incoming customers when the realized service quality is high or low, respectively. Then, $1-f_h$ $(1-f_l)$ is the corresponding probability that the queue length information is concealed from customers when the realized service

quality is high (low). The server then commits to this strategy and announces it to all of the customers, with the goal of maximizing the expected effective arrival rate in his service system.

After the service quality is realized, the corresponding queue-disclosure action is implemented following the pre-announced strategy. Upon observing the server's queue-disclosure action, customers update their beliefs about the service quality according to Bayes' rule. Specifically, when the queue length is revealed, customers assess the service quality to be high with a probability of $P_{H|R}(f_h, f_l) = \frac{\delta_0 f_h}{\delta_0 f_h + (1-\delta_0) f_l}$ and to be low with the complementary probability $1 - P_{H|R}(f_h, f_l)$. Next, they decide whether to join the queue under the assumption that the expected service value is $V_R(f_h, f_l) = P_{H|R}(f_h, f_l)V_h + (1 - P_{H|R}(f_h, f_l))V_l$. According to Naor (1969), customers adopt a threshold policy for joining: they join the queue if and only if the queue length is smaller than some threshold $n_e(f_h, f_l) := \lfloor V_R(f_h, f_l)\mu/\theta \rfloor$, where $\lfloor \cdot \rfloor$ is the floor function. Hence, the queue in equilibrium becomes an $M/M/1/n_e(f_h, f_l)$ system, and the corresponding effective arrival rate, denoted by $\lambda_e^R(f_h, f_l)$, can be calculated as

$$\lambda_e^R(f_h, f_l) = \lambda \left(1 - \frac{\rho^{n_e(f_h, f_l)}}{\sum_{j=0}^{n_e(f_h, f_l)} \rho^j} \right).$$

Similarly, define $P_{H|C}(f_h, f_l)$ and $V_C(f_h, f_l)$ for the case in which the server conceals his queue length. When the queue is concealed, the customers' equilibrium queueing strategy can be represented by the probability of joining the queue (see Edelson and Hildebrand, 1975). This is denoted as $p_e(f_h, f_l)$ and equals 1 if $\lambda < \mu - \theta/V_C(f_h, f_l)$ and $\frac{\mu - \theta/V_C(f_h, f_l)}{\lambda}$ otherwise. The effective arrival rates, denoted as $\lambda_e^C(f_h, f_l)$, are then calculated as λ and $\mu - \theta/V_C(f_h, f_l)$, respectively.

Given the pre-determined queue-disclosure strategy, the probability that the queue is revealed (concealed) is $\delta_0 f_h + (1-\delta_0) f_l$ ($\delta_0(1-f_h) + (1-\delta_0)(1-f_l)$), and the posterior probability of service quality being high is $P_{H|R}(f_h, f_l)$ ($P_{H|C}(f_h, f_l)$). According to Bayes' rule, the expected posterior is equal to the prior, i.e.,

$$\delta_0 = [\delta_0 f_h + (1-\delta_0) f_l] P_{H|R}(f_h, f_l) + [\delta_0(1-f_h) + (1-\delta_0)(1-f_l)] P_{H|C}(f_h, f_l).$$

If a distribution of posterior probabilities satisfies this property, it is called *Bayes-plausible* (see Kamenica and Gentzkow, 2011).

The sequence of events can be summarized as follows. First, the server chooses and commits to a queue-disclosure strategy profile (f_h, f_l). After that, nature determines the service quality and the server makes his queue-disclosure action based on (f_h, f_l). Upon observing the server's disclosure action, customers update their belief about the service quality to be high, represented by $P_{H|R}(f_h, f_l)$ if the queue length is revealed or $P_{H|C}(f_h, f_l)$ if the queue length is concealed. Customers then make their corresponding joining-or-balking decisions. See Figure 1 for an illustration. Backward induction is adopted to derive the game outcome.

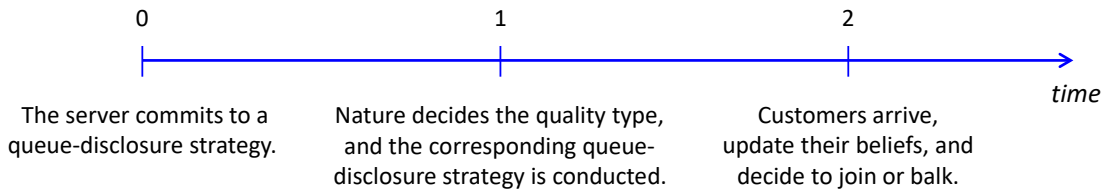

Figure 1: Sequence of events

First, given the queue-disclosure strategy profile (f_h, f_l) and the server's action, we can derive the customers' queueing strategy $(n_e(f_h, f_l), p_e(f_h, f_l))$. We then solve the optimization problem for the server who aims to maximize the expected effective arrival rate:

$$\lambda_e(f_h, f_l) = [\delta_0 f_h + (1-\delta_0) f_l]\lambda_e^R(f_h, f_l) + [\delta_0(1-f_h) + (1-\delta_0)(1-f_l)]\lambda_e^C(f_h, f_l).$$

The server's optimal queue-disclosure strategy is denoted as (f_h^e, f_l^e).

Given the server's queue-disclosure strategy profile, (f_h, f_l), the total utility of all customers under a revealed queue can be derived as

$$u_e^R(f_h, f_l) = \lambda \sum_{j=0}^{n_e(f_h,f_l)-1} p_j^{n_e(f_h,f_l)} \left(V_R(f_h, f_l) - \frac{(j+1)\theta}{\mu} \right),$$

where $p_j^m = \frac{\rho^j}{\sum_{k=0}^m \rho^k}$ $(0 \le j \le m)$. Similarly, the total utility of all customers under a concealed queue can be written as

$$u_e^C(f_h, f_l) = \lambda p_e(f_h, f_l) \left(V_C(f_h, f_l) - \frac{\theta}{\mu - \lambda p_e(f_h, f_l)} \right).$$

Then, the expected total utility across customers can be expressed as

$$u_e(f_h, f_l) = [\delta_0 f_h + (1-\delta_0) f_l]u_e^R(f_h, f_l) + [\delta_0(1-f_h) + (1-\delta_0)(1-f_l)]u_e^C(f_h, f_l).$$

The optimal disclosure strategy may be a *pure* strategy, in which the server commits to fully reveal or conceal his queue length regardless of the service quality level (i.e., f_h and f_l can only be 0 or 1); a *mixed* strategy, in which the server randomizes queue revealing and concealing at both quality levels (i.e., f_h and f_l are both larger than 0 and less than 1); or a *hybrid* strategy, in which the server randomizes queue revealing and concealing at one quality level and fully reveals or conceals the queue length at the other level (i.e., either f_h or f_l is either 0 or 1, and the other is strictly between 0 and 1). We further call the queue disclosure strategy *quality-independent* if the probability that the server reveals the queue length is the same for both high and low service quality levels and *quality-dependent* if the probability differs by the service quality level. Clearly, the

quality-independent disclosure strategy conveys no information about service quality, and hence the posterior equals the prior. Only a quality-dependent disclosure strategy conveys some information about service quality to customers.

4 Optimal Queue Disclosure Strategy

In this section, we analyze the server's optimal queue disclosure strategy. First, we reformulate the server's decision problem as a nonlinear program, the optimal solution of which can be derived geometrically through convex combination. We then investigate the effect of market size on system performance.

4.1 Geometric Approach

We can directly maximize the server's expected effective arrival rate by considering the disclosure probabilities (f_h, f_l) as decision variables. However, this approach does not yield closed-form solutions and thus cannot provide useful insights. Below, we consider the problem from another angle. We first demonstrate a one-to-one correspondence between the server's queue disclosure strategy and the Bayes-plausible distribution of posteriors. We then transform the server's optimization problem into a new problem on finding the best Bayes-plausible distribution of posteriors. We then provide a geometric approach to derive the optimal disclosure strategy.

In Section 3, we show that the server's queue disclosure strategy yields a unique Bayes-plausible distribution of posterior beliefs. Conversely, any Bayes-plausible distribution of posterior beliefs corresponds to a unique queue disclosure strategy. The details are as follows. Suppose that customers observe a revealed queue with a probability of p^R and a concealed queue with a probability of $p^C = 1 - p^R$. The posterior belief about the service quality to be high conditional on a revealed queue is $p_{H|R}$, and the effective arrival rate is a function of this posterior belief, denoted by $\lambda_e^R(p_{H|R})$. Similarly, the posterior belief about the service quality to be high conditional on a concealed queue is $p_{H|C}$, and the corresponding effective arrival rate is a function of this belief, denoted by $\lambda_e^C(p_{H|C})$. We then have the following proposition.

Proposition 1. *Consider a prior δ_0 and two posteriors, $p_{H|R}$ with probability p^R when the queue length is revealed and $p_{H|C}$ with probability p^C when the queue length is concealed. If such a distribution of posteriors is Bayes-plausible (i.e., $\delta_0 = p^R p_{H|R} + p^C p_{H|C}$), it can be induced by a queue disclosure strategy with $f_h = p^R p_{H|R}/\delta_0$ and $f_l = p^R(1 - p_{H|R})/(1 - \delta_0)$.*

Based on Proposition 1, we can transform our search for the optimal disclosure strategy into a search for the best Bayes-plausible distribution of posterior beliefs. Mathematically, we can rewrite

the server's effective arrival rate maximization problem as follows:

$$\begin{aligned} \lambda_e(f_h^e, f_l^e) \quad = \quad & \max_{p^R,\ p^C,\ p_{H|R},\ p_{H|C}} p^R \lambda_e^R(p_{H|R}) + p^C \lambda_e^C(p_{H|C}) \\ s.t. \qquad & p^R + p^C = 1 \\ & \delta_0 = p^R p_{H|R} + p^C p_{H|C} \\ & 0 \le p^R,\ p^C,\ p_{H|R},\ p_{H|C} \le 1. \end{aligned} \tag{1}$$

This optimization problem can be solved through a geometric approach, which we now describe in detail. Let δ represent the parameter of the posterior belief (i.e., the probability of service quality being high). The effective arrival rate, which is conditional on a revealed or a concealed queue, is a function of the expected service quality, which is determined by the customers' posterior belief δ. Therefore, we can express the effective arrival rates as functions of the posterior belief δ. Now, consider the two effective arrival rate functions, $\lambda_e^R(\delta)$ and $\lambda_e^C(\delta)$, in the domain $\delta \in [0, 1]$. Recall that $p^R + p^C = 1$. When p^R changes from 0 to 1, the value of $p^R \lambda_e^R(p_{H|R}) + p^C \lambda_e^C(p_{H|C})$ lies on the line segment connecting the two points $(p_{H|R}, \lambda_e^R(p_{H|R}))$ and $(p_{H|C}, \lambda_e^C(p_{H|C}))$. The point where this line segment crosses the vertical line $\delta = \delta_0$ satisfies the Bayes plausibility requirement (1). Therefore, to find the optimal solution, we only need to consider all of the segments that connect a point on the function curve of $\lambda_e^R(\delta)$ and a point on the function curve of $\lambda_e^C(\delta)$. The highest crossing point of all possible line segments with the vertical line $\delta = \delta_0$ represents the maximal effective arrival rate that can be achieved through the server's queue disclosure strategy.

To facilitate derivation of the structural properties of this reformulated optimization problem, we first provide the following lemma on the shapes of the two effective arrival rate functions.

Lemma 1. *The two effective arrival rate functions, $\lambda_e^R(\delta)$ and $\lambda_e^C(\delta)$, exhibit the following properties:*

(i) $\lambda_e^R(\delta)$ is a piecewise constant function with some upward jumps as δ increases from 0 to 1;

(ii) $\lambda_e^C(\delta)$ is concave and nondecreasing in δ.

The shapes of the two effective arrival rate functions can be used to derive the optimal queue disclosure strategy. For the sake of analysis, we further define the point set

$$co(\lambda_e^R(\cdot), \lambda_e^C(\cdot)) := \{\alpha(\delta_1, \lambda_e^R(\delta_1)) + (1-\alpha)(\delta_2, \lambda_e^C(\delta_2)) | 0 \le \alpha, \delta_1, \delta_2 \le 1\},$$

which contains all the convex combinations of one point $(\delta_1, \lambda_e^R(\delta_1))$ on the function $\lambda_e^R(\cdot)$ and another point $(\delta_2, \lambda_e^C(\delta_2))$ on the function $\lambda_e^C(\cdot)$, where $0 \le \delta_1, \delta_2 \le 1$. The significance of constructing $co(\lambda_e^R(\cdot), \lambda_e^C(\cdot))$ is demonstrated in the following proposition.

Proposition 2. *Given a prior belief δ_0, there exists a queue disclosure strategy (f_h, f_l) that results in an expected effective arrival rate $\lambda_e(f_h, f_l)$ if and only if $(\delta_0, \lambda_e(f_h, f_l)) \in co(\lambda_e^R(\cdot), \lambda_e^C(\cdot))$.*

Proposition 2 ensures that the maximal effective arrival rate can be searched only in the set $co(\lambda_e^R(\cdot), \lambda_e^C(\cdot))$. Define

$$\Lambda_e(\delta) := \max\{\Lambda | (\delta, \Lambda) \in co(\lambda_e^R(\cdot), \lambda_e^C(\cdot))\}. \tag{2}$$

Then, function $\Lambda_e(\delta)$ with $\delta \in [0, 1]$ is the upper envelope of the set $co(\lambda_e^R(\cdot), \lambda_e^C(\cdot))$.

Based on Proposition 2, we have the following conclusion on the optimal queue disclosure strategy.

Proposition 3. *Given the prior δ_0, the maximal effective arrival rate under the optimal queue disclosure strategy is $\Lambda_e(\delta_0)$.*

Proposition 3 indicates that a pre-committed queue-disclosure strategy benefits the server at the given prior δ_0 only if $\Lambda_e(\delta_0) > \max\left\{\lambda_e^R(\delta_0), \lambda_e^C(\delta_0)\right\}$. A similar upper envelope is provided in Kamenica and Gentzkow (2011). In that work, however, different signals correspond to the same value function of the sender, and thus the upper envelope is formed through the concavification of that value function. In contrast, in our work, the queue-disclosure actions– revealing and concealing the queue length– are the signals. These two signals correspond to two different value functions. Under the Bayes plausibility condition, the upper envelope is formed through the convex combination of these two value functions. We now illustrate the aforementioned geometric approach in the following example.

Example 2. (Illustration of the Upper Envelope) *Consider the parameter values to be $V_h = 2$, $V_l = 1$, $\mu = 1.1$, $\theta = 1$ and $\lambda = 0.6$. The dashed curve in Figure 2 represents the effective arrival rate function $\lambda_e^C(\delta)$, and the dotted piecewise flat line represents the effective arrival rate function $\lambda_e^R(\delta)$. The solid line connecting the two points $(0, \lambda_e^R(0))$ and $(1, \lambda_e^C(1))$ is the upper envelope formed by all of the segments connecting two arbitrary points on these two effective arrival rate functions. The first point $(0, \lambda_e^R(0))$ represents the effective arrival rate of a revealed queue with a posterior belief $p_{H|R} = 0$, and the second point $(1, \lambda_e^C(1))$ represents the effective arrival rate of a concealed queue with a posterior belief $p_{H|C} = 1$. Given any prior belief δ_0 (e.g., $\delta_0 = 0.3$), we can recover the probability p^R by solving the Bayes plausibility condition $p^R * 0 + (1 - p^R) * 1 = 0.3$, which yields $p^R = 0.7$. Then, according to Proposition 1, we can recover the optimal queue disclosure strategy as follows: $f_h = p^R p_{H|R}/\delta_0 = 0$ and $f_l = p^R(1 - p_{H|R})/(1 - \delta_0) = 1$. One can easily check that for any prior belief $\delta_0 \in (0, 1)$, the optimal queue disclosure strategy is to always conceal the queue length when the realized service quality is high but to always reveal it when the realized service*

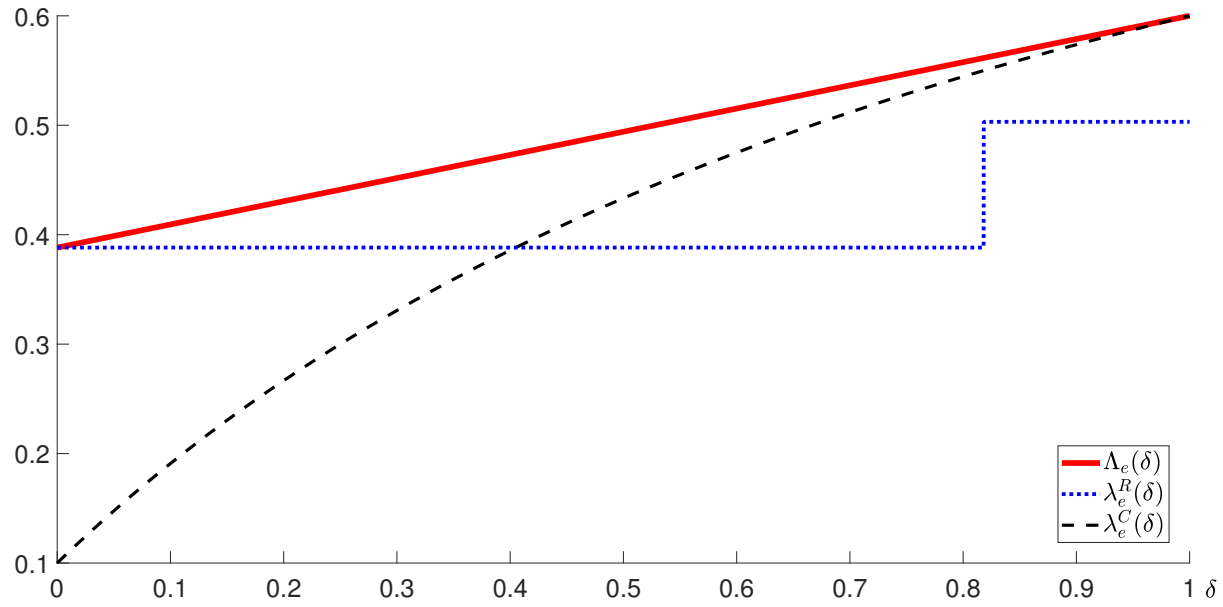

Figure 2: The upper envelope Λ_e: $V_h = 2$, $V_l = 1$, $\mu = 1.1$, $\theta = 1$ and $\lambda = 0.6$

quality is low, i.e., $(f_h^e, f_l^e) = (0, 1)$. *That is, the server's optimal queue-disclosure strategy is pure and quality-dependent and thus fully conveys quality information to customers.*

Example 2 shows that a pure disclosure strategy can be the server's optimal strategy. Below, we demonstrate that a hybrid disclosure strategy works best for the server under the setting given in Example 1 (stated in the Introduction).

Example 3. (Illustration of Example 1 via Geometric Approach) *The motivating Example 1 is illustrated in Figure 3. In this example, the effective arrival rate function of the concealed queue reaches a linear plateau at* $d_{12} = 0.6667$. *The solid line represents the upper envelope formed by all of the segments connecting two arbitrary points of the two effective arrival rate functions. Given the prior belief* $\delta_0 = 0.33$, *the maximal effective arrival rate is located on the segment connecting the two points* $(0, \lambda_e^R(0))$ *and* $(d_{12}, \lambda_e^C(d_{12}))$. *The first point,* $(0, \lambda_e^R(0))$, *represents the effective arrival rate in a revealed queue with a posterior belief* $p_{H|R} = 0$, *and the second point,* $(d_{12}, \lambda_e^C(d_{12}))$, *represents the effective arrival rate in a concealed queue with a posterior belief* $p_{H|C} = 0.6667$. *Given* $\delta_0 = 0.33$, *we can recover the probability* p^R *by solving the Bayes plausibility condition* $p^R * 0 + (1 - p^R) * 0.6667 = 0.33$, *which yields* $p^R = 0.5050$. *Then, according to Proposition 1, we can recover the optimal queue disclosure strategy as follows:* $f_h = p^R p_{H|R}/\delta_0 = 0$ *and* $f_l = p^R(1 - p_{H|R})/(1 - \delta_0) = 0.7537$. *This hybrid strategy conveys partial information about service quality to customers. By checking the graph of the upper envelope, we can see that for any prior belief* $\delta_0 \in (0, d_{12})$, $\Lambda_e(\delta_0)$ *is located above the two effective arrival rate functions, and the corresponding hybrid queue-disclosure strategy is beneficial to the server.*

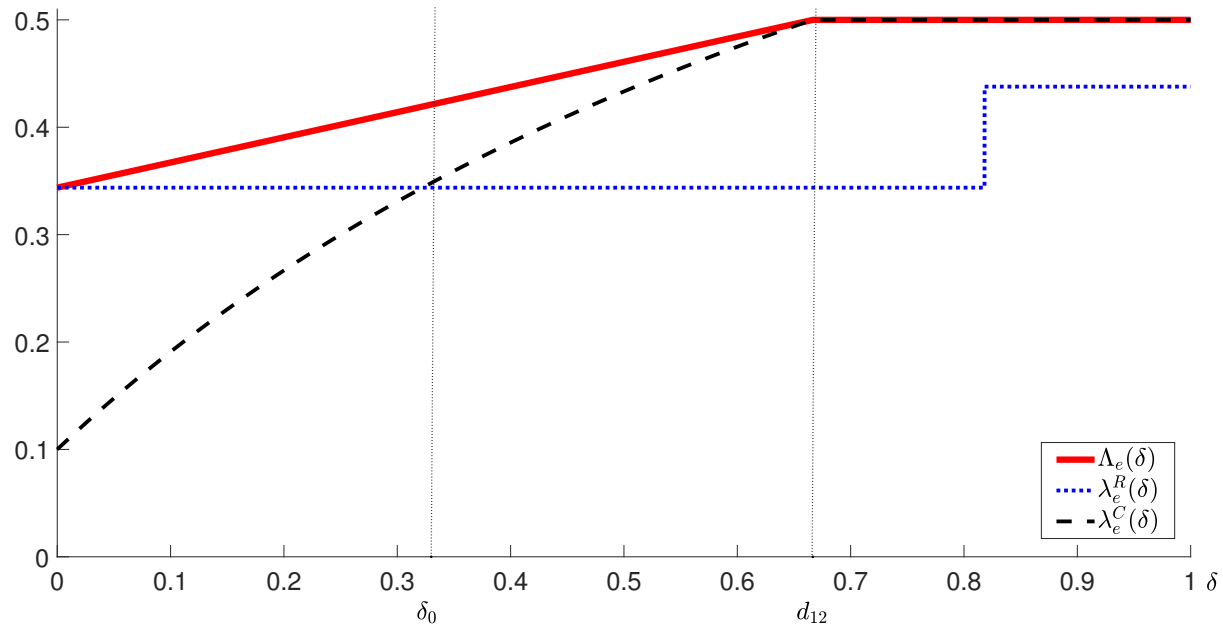

Figure 3: The upper envelope Λ_e: $V_h = 2$, $V_l = 1$, $\mu = 1.1$, $\theta = 1$ and $\lambda = 0.5$

We note that although the maximal effective arrival rate is unique for a given prior belief δ_0, the corresponding optimal queue-disclosure strategy is not necessarily unique. A point on the upper envelope may correspond to multiple pairs of posteriors whose distribution is Bayes-plausible. Let us revisit Example 2 and the upper envelope plotted in Figure 2. We retain the parameter values $V_h = 2$, $V_l = 1$, $\mu = 1.1$ and $\theta = 1$ but change the value of λ from 0.6 to 0.7160. In this setting, the up-jumping point of $\lambda_e^R(\delta)$, $(0.8182, 0.5698)$, is located on the upper envelope $\Lambda_e(\delta)$, a segment connecting the two points $(0, \lambda_e^R(0))$ and $(1, \lambda_e^C(1))$, where $\lambda_e^R(0) = 0.4337$ and $\lambda_e^C(1) = 0.6000$. When the prior is $\delta_0 = 0.8182$, we can obtain the following two optimal queue-disclosure strategies: $(f_h^e, f_l^e) = (0, 1)$ or $(1, 1)$.

Based on the above geometric approach, we further obtain the following lemma.

Lemma 2. *For any prior δ_0, the optimal queue disclosure strategy is pure and quality-independent in the following two situations:*

(i) $(f_h^e, f_l^e) = (0, 0)$ *if* $\lambda_e^C(\delta) > \lambda_e^R(\delta)$ *for all* $\delta \in [0, 1]$.

(ii) $(f_h^e, f_l^e) = (1, 1)$ *if* $\lambda_e^R(\delta)$ *is a constant function (i.e., a horizontal line) and* $\lambda_e^R(\delta) \geq \lambda_e^C(\delta)$ *for all* $\delta \in [0, 1]$.

The first statement of Lemma 2 requires that the effective arrival rate function of a concealed queue be located above that of a revealed queue. By considering its concavity property (see Lemma 1), we conclude that the upper envelope function coincides with the effective arrival rate

function of a concealed queue. Hence, the optimal queue disclosure strategy is to always conceal the queue, regardless of the realized service quality. The second statement of Lemma 2 provides a sufficient condition under which the optimal queue disclosure strategy is to always reveal the queue, regardless of the realized service quality. Note that this condition requires the effective arrival rate function of a revealed queue not only to be located above that of a concealed queue but also to be a constant function, such that no jumps in this function occur in the whole domain $\delta \in [0, 1]$.

4.2 Effect of Market Size

In this section, we set a fixed prior belief δ_0 and explore the effect of market size (i.e., the potential arrival rate) λ on the server's optimal queue disclosure strategy.

When the service quality is certain, the effect of market size on the delay announcement strategy is well studied in the literature. According to Hassin (1986) and Chen and Frank (2004), when the market size λ is below a threshold value, concealing the queue benefits the server; otherwise, revealing the queue is preferred. Moreover, when λ is very small, customers 'all join' in an unobservable queue setting while some balk in an observable queue setting. Hence, concealing queue length information is the better option for servers when λ is very small. As λ becomes sufficiently large, the effective arrival rate becomes constant for unobservable queues because customers' joining utility is now zero and no more customers want to join the queue. However, in an observable queue setting, the queue becomes stochastically longer as λ increases, and the effective arrival rate strictly increases because the number of customers who observe a short queue and join continues to increase. Therefore, the server benefits more from revealing the queue when λ is very large. In our work, the aforementioned results and insights still hold in sufficiently small- and large-sized markets under certain conditions, as implied by Lemma 2. We now formally show that these results also hold for our optimal queue disclosure strategy.

Proposition 4. *Given any prior δ_0, the optimal queue disclosure strategy, (f_h^e, f_l^e), satisfies the following two properties:*

(i) If the potential arrival rate $\lambda < \mu - \frac{\theta}{\delta_0 V_h + (1-\delta_0) V_l}$, then the server's optimal strategy is to always conceal the queue; that is, $(f_h^e, f_l^e) = (0, 0)$.

(ii) There exists a threshold, denoted by $\bar{\lambda}^e$ (which is greater than $\mu - \frac{\theta}{\delta_0 V_h + (1-\delta_0) V_l}$), such that if $\lambda > \bar{\lambda}^e$,[1] the server's optimal strategy is to always reveal the queue; that is, $(f_h^e, f_l^e) = (1, 1)$.

Indeed, when the market size is very small, all customers join the concealed queue regardless of the level of service quality, and thus concealing the queue is the server's optimal strategy. Similarly,

[1]The definition of $\bar{\lambda}^e$ can be found in the proof of Proposition 4.

when the market size is very large, revealing the queue is the optimal strategy. However, when the market size λ is medium, the situation becomes tricky, and the optimal disclosure strategy depends on the tradeoff between the value of informing customers about the queue length and that of providing partial quality information. We use the following numerical example to illustrate this tradeoff.

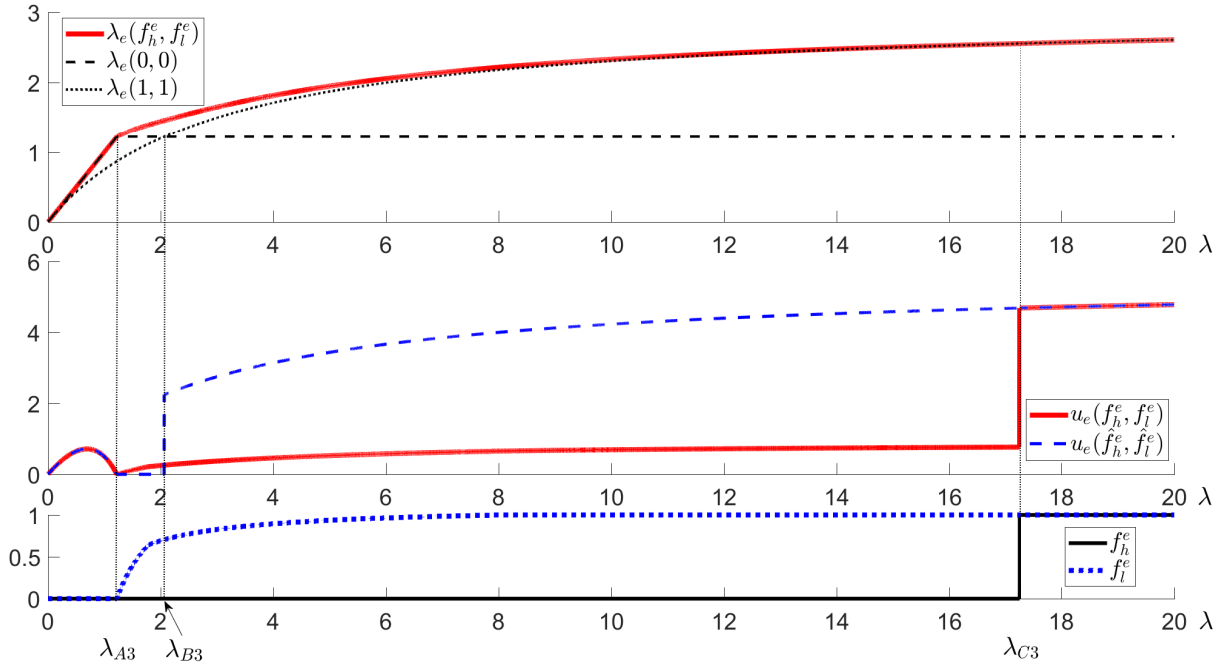

Figure 4: Effect of market size on the optimal queue disclosure strategy, maximal effective arrival rate and customers' total utility: $V_h = 18$, $V_l = 3$, $\mu = 3$, $\theta = 8$ and $\delta_0 = 0.1$

Example 4. (Sensitivity Analysis: Effect of Market Size λ on the Server's Optimal Queue Disclosure Strategy and System Performance) *Consider the parameter values $V_h = 18$, $V_l = 3$, $\mu = 3$, $\theta = 8$ and $\delta_0 = 0.1$. There are three key market size thresholds, as shown in Figure 4: $\lambda_{A3} = 1.2222$, $\lambda_{B3} = 2.0621$ and $\lambda_{C3} = 17.2492$. In that figure, the bottom subplot depicts the server's optimal queue disclosure strategy (f_h^e, f_l^e) as a function of λ; the middle subplot shows the customers' total utility; and the upper subplot depicts the maximal effective arrival rate that can be achieved by adopting the pre-committed optimal queue disclosure strategy (f_h^e, f_l^e).*

Figure 4 shows that when the market size is small ($\lambda < \lambda_{A3}$), fully concealing the queue (i.e., $(f_h^e, f_l^e) = (0,0)$) is the dominant strategy because in such a situation, all customers join the concealed queue. When the market size reaches the threshold λ_{A3}, balking becomes possible because customers' expected joining utility is now reduced to zero. As the market size further increases and becomes larger than λ_{A3}, the server continues to conceal the queue when the realized service quality is high but begins to randomize actions to conceal and reveal the queue when the realized service

quality is low; here, the probability of revealing the queue increases with the market size λ. On the one hand, such randomization can strengthen customers' belief that the service quality is high when they observe that the queue length is concealed; on the other hand, this approach provides a chance for the customers to see a revealed queue, from which they then infer that the service quality is low. Still, the increase in the customers' effective arrival rate in the concealed case surpasses the reduction in the customers' effective arrival rate in the revealed case, thereby benefiting the server. When the market size reaches and increases beyond λ_{C3}, it is no longer beneficial to conceal the queue if the realized service quality is high, as a revealed queue induces more customers to join in this condition. Accordingly, fully revealing the queue is the dominant strategy, that is, $(f_h^e, f_l^e) = (1, 1)$.

Consequently, in a small-sized market ($\lambda \in (0, \lambda_{A3})$), the maximal effective arrival rate coincides with that of a fully concealed queue, whereas that in a large-sized market ($\lambda \in (\lambda_{C3}, +\infty)$) coincides with that of a fully revealed queue, as shown in the upper subplot of Figure 4. However, in a medium-sized market ($\lambda \in (\lambda_{A3}, \lambda_{C3})$), the maximal effective arrival rate is strictly larger than that of either a fully revealed or fully concealed queue. The effective arrival rate difference between them can be used to determine the value of using a quality-dependent disclosure strategy.

We also derive the optimal quality-independent queue disclosure strategy, denoted by $(\hat{f}_h^e, \hat{f}_l^e)$, and the corresponding total customer utility to understand the effect of the quality-dependent queue disclosure strategy on customers. The middle subplot of Figure 4 shows that compared with the quality-independent queue disclosure strategy, our quality-dependent queue disclosure strategy can improve customers' total utility only when the market size λ falls within a relatively small range: $(\lambda_{A3}, \lambda_{B3})$. However, for a relatively large market size within the range of $\lambda \in (\lambda_{B3}, \lambda_{C3})$, customers' total utility is smaller under our optimal quality-dependent queue disclosure strategy than under the optimal quality-independent queue disclosure strategy. Therefore, although the pre-committed queue disclosure strategy can be used to attract more customers to join the service system, it does not necessarily benefit them.

Note that in Example 4, when the market size λ falls within the range of $(\lambda_{A3}, \lambda_{C3})$, a hybrid equilibrium is sustained in which the server randomizes concealing and revealing the queue only when the realized service quality is low. We also conduct other numerical examples and find that the equilibrium may also be fully mixed, such that the server randomizes queue concealment and revelation at both high and low levels of service quality.

5 Social Planner

In the previous section, we consider a profit-maximizing server and study his optimal queue disclosure strategy. In reality, however, servers can be social planners whose aim is to maximize overall

social welfare. We now extend our commitment game to this setting and show that our geometric approach is robust and provides new insights into the optimal queue-disclosure strategy.

In our setting, social welfare is defined as the sum of customers' utilities, represented by

$$u_e(f_h, f_l) = [\delta_0 f_h + (1-\delta_0) f_l] u_e^R(f_h, f_l) + [\delta_0(1-f_h) + (1-\delta_0)(1-f_l)] u_e^C(f_h, f_l),$$

where detailed expressions of $u_e^R(f_h, f_l)$ and $u_e^C(f_h, f_l)$ are provided in Section 3. To analyze the social planner's optimal queue disclosure strategy, we first investigate the geometry underlying the functions involved and obtain the following results regarding the shapes of two utility functions, $u_e^R(\cdot)$ and $u_e^C(\cdot)$.

Lemma 3. *The two utility functions, $u_e^R(\delta)$ and $u_e^C(\delta)$, exhibit the following properties:*

(i) $u_e^R(\delta)$ is a piecewise increasing linear function with some downward jumps as δ increases from 0 to 1;

(ii) if $\lambda \geq \mu - \frac{\theta}{V_h}$, then $u_e^C(\delta)$ is equal to 0 for all $\delta \in [0,1]$; if $\lambda \leq \mu - \frac{\theta}{V_l}$, then $u_e^C(\delta)$ is linear and increasing in δ for $\delta \in [0,1]$; otherwise, $u_e^C(\delta)$ is equal to 0 for $\delta \in \left[0, \frac{\theta/(\mu-\lambda)-V_l}{V_h-V_l}\right]$ and linear and increasing in δ for $\delta \in \left(\frac{\theta/(\mu-\lambda)-V_l}{V_h-V_l}, 1\right]$;

(iii) $u_e^R(\delta) > u_e^C(\delta)$ for all $\delta \in [0,1]$.

The third statement of Lemma 3 shows that from the perspective of welfare maximization, revealing the queue is always better than concealing it (see also Hassin and Haviv, 2003; and Hassin and Roet-Green, 2017). This is intuitive, as revealing the queue can help customers to make an informed decision. Considering the optimal queue disclosure strategy, however, this classical result no longer holds. We show here that concealing the queue length with a strictly positive probability is socially desirable under some situations.

Using the geometric approach, we construct the convex-combination point set as

$$co(u_e^R(\cdot), u_e^C(\cdot)) := \{\alpha(\delta_1, u_e^R(\delta_1)) + (1-\alpha)(\delta_2, u_e^C(\delta_2)) | 0 \leq \alpha, \delta_1, \delta_2 \leq 1\},$$

and the upper envelope as

$$U_e(\delta) := \max\{U | (\delta, U) \in co(u_e^R(\cdot), u_e^C(\cdot))\}.$$

The social planner's optimal queue disclosure strategy, $(\tilde{f}_h^e, \tilde{f}_l^e)$, can thus be derived accordingly.

Example 5. (Illustration of the Social Planner's Optimal Queue Disclosure Strategy) *Consider a setting in which $V_h = 2$, $V_l = 1$, $\mu = 1.1$, $\theta = 1$ and $\lambda = 10$. The dotted piecewise*

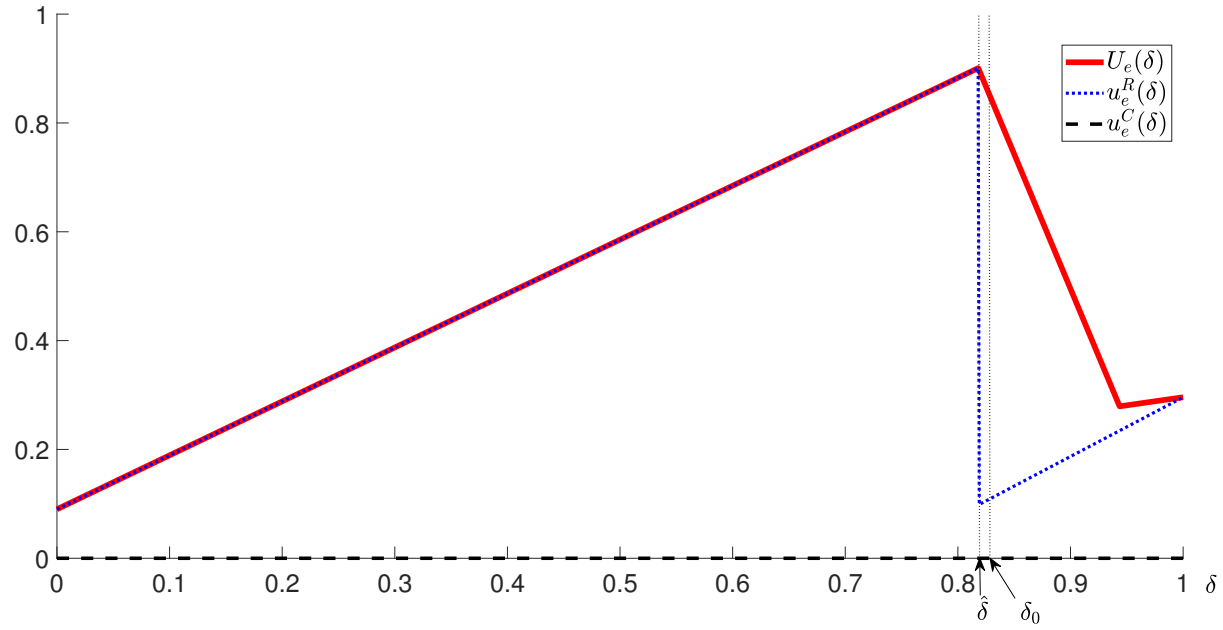

Figure 5: The upper envelope U_e: $V_h = 2$, $V_l = 1$, $\mu = 1.1$, $\theta = 1$ and $\lambda = 10$

increasing line in Figure 5 represents the utility function $u_e^R(\delta)$*, the flat line represents the utility function* $u_e^C(\delta)$*, and the solid line is the upper envelope formed by all segments connecting two arbitrary points on these two utility functions. Given the prior belief* $\delta_0 = 0.83$*, the maximal effective arrival rate is located on the segment connecting the two points* $(\hat{\delta}^-, u_e^R(\hat{\delta}^-))$ *and* $(1, u_e^C(1))$*, where* $\hat{\delta}$ *is the down-jumping point of the utility function* $u_e^R(\delta)$*,* $\hat{\delta}^- = \lim_{\delta \to \hat{\delta}, \delta < \hat{\delta}} \delta$ *and* $u_e^R(\hat{\delta}^-)$ *is the left-hand limit of* $u_e^R(\cdot)$ *at the point* $\hat{\delta}$*. The first point,* $(\hat{\delta}^-, u_e^R(\hat{\delta}^-))$*, represents the effective arrival rate in a revealed queue with a posterior belief* $p_{H|R} = 0.8181$*, and the second point,* $(1, u_e^C(1))$*, represents the effective arrival rate in a concealed queue with a posterior belief* $p_{H|C} = 1$*. Given* $\delta_0 = 0.83$*, we can recover the probability* p^R *by solving the Bayes plausibility condition* $p^R * 0.8181 + (1 - p^R) * 1 = 0.83$*, which yields* $p^R = 0.9346$*. Then, according to Proposition 1, we can obtain the optimal queue disclosure strategy as* $\tilde{f}_h^e = p^R p_{H|R}/\delta_0 = 0.9212$ *and* $\tilde{f}_l^e = p^R(1 - p_{H|R})/(1 - \delta_0) = 1.0000$*. Clearly, this optimal strategy is hybrid. Figure 5 indicates that under the prior* $\delta_0 = 0.83$*, the expected total utility under the optimal queue disclosure strategy (i.e.,* $U_e(\delta_0) = 0.8419$*) improves by* 659% *over the one under the "always revealing" strategy (i.e.,* $u_e^R(\delta_0) = 0.1109$*).*

Let $\hat{\Delta}$ denote the set of values of δ where $u_e^R(\delta)$ jumps down (or specifically, the term $\mu[\delta V_h + (1-\delta)V_l]/\theta$ takes integer values). Noticing that $\delta \in [0, 1]$, we can formally define

$$\hat{\Delta} := \left\{ \delta | \delta = \frac{k\theta - \mu V_l}{\mu(V_h - V_l)}, k = \lceil \mu V_l/\theta \rceil, \cdots, \lfloor \mu V_h/\theta \rfloor \right\},$$

where $\lceil \cdot \rceil$ and $\lfloor \cdot \rfloor$ are the ceiling and floor functions, respectively. Then, consider any prior $\delta_0 \in$

$[\hat{\delta}, \hat{\delta}+\epsilon)$, where $\hat{\delta} \in \hat{\Delta}$, ϵ is sufficiently small and $\hat{\delta}+\epsilon < 1$. For the point $(\delta_0, \bar{u})$ on the segment connecting two points $(\hat{\delta}^-, u_e^R(\hat{\delta}^-))$ and $(1, u_e^C(1))$, $\bar{u}$ is strictly larger than $u_e^R(\delta_0)$, which implies that a greater expected utility must be achieved with the optimal queue disclosure strategy than with the "always revealing" strategy. Also, $\hat{\delta}$ is independent of the market size λ. In summary, we reach the following conclusion.

Proposition 5. *Given any prior $\delta_0 \in [\hat{\delta}, \hat{\delta}+\epsilon)$ where $\hat{\delta} \in \hat{\Delta}$, ϵ is sufficiently small and $\hat{\delta}+\epsilon < 1$, the expected total utility is greater with the optimal queue disclosure strategy than with the "always revealing" strategy on the whole range of market size; that is, $U_e(\delta_0) > u_e^R(\delta_0)$ for all $\lambda \in (0, +\infty)$.*

Proposition 5 implies that full disclosure is not necessarily socially desired. Combining it with the third statement of Lemma 3, we can conclude that it can be optimal for the social planner to randomize queue concealment and revelation, regardless of the market size. This conclusion echoes those reached by Cui and Veeraraghavan (2016), Hu et al. (2018) and Li et al. (2020). According to Naor (1969), tolls/taxes can be levied in queueing systems to control arrivals with the intention of improving welfare. In Cui and Veeraraghavan (2016), the lack of information acts as an information tax that deters admission, leading to improved welfare. A similar rationale holds here.

In contrast to the profit-maximizing case in which the server's optimal queue disclosure strategy is "always concealing" ("always revealing") when the market size λ is sufficiently small (large), a social planner's optimal queue disclosure strategy may be quality-dependent on the whole range of λ, as illustrated in the following example.

Example 6. (Sensitivity Analysis: Effect of Market Size λ on Social Planner's Optimal Queue Disclosure Strategy and System Performance) *Consider the setting in Example 5, in which the prior belief δ_0 is very close to the down-jumping point $\hat{\delta}$. Figure 6 indicates that across the whole market size range ($\lambda \in (0, +\infty)$), the optimal queue disclosure strategy, $(\tilde{f}_h^e, \tilde{f}_l^e)$, achieves a strictly larger expected total utility than the "always revealing" strategy, $(\tilde{f}_h, \tilde{f}_l) = (1, 1)$.*

In the profit-maximizing case, more arrivals are always preferred. In the welfare-maximization case, however, it may be socially desirable to persuade fewer customers to join, as an overly crowded system can reduce customers' overall utility. To discourage some customers from joining, the social planner should convince them that the service quality may be low. In Figure 6, for $0 < \lambda < \lambda_D (= 0.1000)$, the optimal queue disclosure strategy is pure, with $(\tilde{f}_h^e, \tilde{f}_l^e) = (0, 1)$; this strategy indirectly provides full information on the quality type. In this case, as the market size λ is very small, the expected total utility from joining a concealed queue is strictly positive when the posterior belief $p_{H|C}$ is 1, which is even larger than expected total utility from joining a revealed queue with the prior

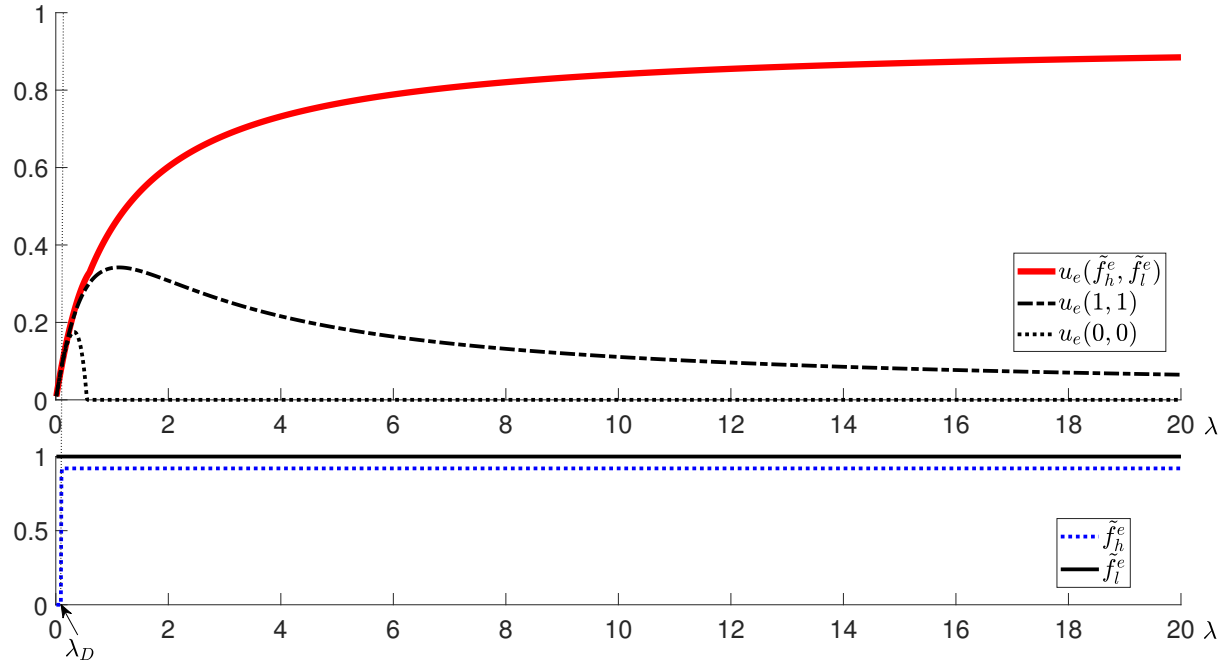

Figure 6: Effect of market size on a social planner's optimal queue disclosure strategy and customers' total utility: $V_h = 2$, $V_l = 1$, $\mu = 1.1$, $\theta = 1$ and $\delta_0 = 0.83$

belief $\delta_0 = 0.83$. *This increase provides an incentive for the social planner to conceal the queue when the service quality level is high. To achieve an increase in overall utility, the server should reveal the queue when the service quality level is low. For* $\lambda_D \leq \lambda < +\infty$, *the customer utility in a concealed queue becomes small, and the total utility from a revealed queue with the prior* $\delta_0 = 0.83$ *is also relatively small. In this case, the social planner should randomize revealing and concealing the queue to weaken customers' belief regarding a high service quality level (i.e.,* $p_{H|R}$*). For example, under the special case of* $\lambda = 10$ *in the previous Example 5, the optimal queue disclosure strategy is* $(\tilde{f}_h^e, \tilde{f}_l^e) = (0.9212, 1.0000)$. *Under such a strategy, customers' belief about a high service quality level after seeing a revealed queue reaches a value of* $0.8181 (< \delta_0)$, *which decreases the maximal queue length from* 2 *to* 1 *and achieves a higher expected total utility than that obtained with the "always revealing" strategy.*

6 Concluding Remarks

The service quality provided in some systems is generally uncertain. In this work, we examine a situation in which the server can design a queue-disclosure strategy that links queue concealment and revelation with the realized level of service quality and commits to it before the service quality is realized. We demonstrate that the commitment to such a disclosure strategy increases the server's ability to attract more customers to join the service system.

We transfer our search for the optimal queue disclosure strategy to an equivalent search for the optimal Bayes-plausible distribution of posterior beliefs. Based on our reformulated optimization problem, we then provide a geometric approach to obtain this optimal strategy. We show that as long as the upper envelope of all convex combinations of one point from the effective arrival rate function of a concealed queue and another point from that of a revealed queue is located above these two functions, a properly designed queue disclosure strategy (which might involve randomization) can be implemented to help attract more customers to join the service system. We also investigate the effect of market size on the server's optimal queue disclosure strategy. We show that it is always in the server's best interest to conceal the queue in a very small-sized market and reveal it in a very large-sized market. Through the numerical study, we find that in a medium-sized market, it is often optimal for the server to adopt a quality-dependent queue-disclosure strategy to either fully or partially convey the service quality information to customers. We then extend our analysis to a setting in which the server is a welfare-maximizing social planner. We show that the geometric approach can still be easily applied to this situation and that it may be beneficial for the social planner to randomize revealing and concealing the queue, regardless of market size. This result is in sharp contrast to the classical literature (see, e.g., Hassin and Haviv, 2003; and Hassin and Roet-Green, 2017), which states that it is always socially optimal to reveal the queue.

Our work demonstrates that our quality-dependent queue-disclosure strategy can be used to better match the realized quality level with queue-disclosure actions and fully or partially convey quality information to customers. We further present an intuitive geometric approach to the identification of such an optimal strategy with the objective of arrival-rate maximization or welfare maximization. Admittedly, our model has some limitations. First, we restrict the signal to a binary choice of concealing or revealing the queue. Under this assumption of a binary choice, the effective arrival rates can be easily calculated. It would be interesting to extend our approach to other types of delay announcements, such as informing customers of the exact waiting time (Guo and Zipkin, 2007) or announcing the waiting time of the last customer to enter service (Ibrahim et al., 2017). Second, our information disclosure scheme is designed based on a one-dimension uncertain state (i.e., quality type). It would be interesting to determine the optimal persuasion mechanism based on two-dimensional uncertain states (i.e., quality type and queue length). Third, we assume that customers cannot obtain quality information from the customers who have been served. Nowadays, with advanced information technology, it is quite convenient for customers to share the information about the realized service quality through online reviews or social networks. It would be interesting to incorporate consumer-generated quality information (e.g., Yu et al., 2016) into our model setting. Despite such limitations, we hope that our work can serve as a stepping stone for further studies on the combination of information disclosure and Bayesian persuasion in queueing systems.

Finally, our model does not fit a setting in which the server does not pre-commit to any strategy and decides to reveal or conceal the queue after the service quality is realized. In such a setting, the server might communicate vague signals about the realized quality type to customers, and the equilibrium between the server and customers could thus be modeled as a cheap talk game. It would be interesting to study the equilibrium of such a game and compare the corresponding equilibrium behaviors with our results from this work.

Acknowledgments

The authors gratefully thank the departmental editor (Prof. Michael Pinedo), an anonymous senior editor, two anonymous referees, and Prof. Refael Hassin for their very helpful comments and suggestions. The first author Pengfei Guo acknowledges the financial support by the Research Grants Council of Hong Kong (No. 15502820). Research of the second author Moshe Haviv was funded by Israel Science Foundation grant no. 1512/19. The corresponding author Zhenwei Luo's work was supported by the Internal Start-up Fund of The Hong Kong Polytechnic University (Project ID: P0039035). The fourth author Yulan Wang acknowledges the the financial support from the Research Grants Council of Hong Kong (RGC Reference Number: 15505019). All authors contributed equally to the work.

References

Aksin, Z., Armony, M. and V. Mehrotra. 2007. The modern call-center: A multi-disciplinary perspective on operations management research. *Production and Operations Management* 16(6) 665-688.

Allon, G., Bassamboo, A. and I. Gurvich. 2011. "We will be right with you": Managing customer expectations with vague promises and cheap talk. *Operations Research* 59(6) 1382-1394.

Armony, M. and C. Maglaras. 2004a. Contact centers with a call-back option and real-time delay information. *Operations Research* 52(4) 527-545.

Armony, M. and C. Maglaras. 2004b. On customer contact centers with a call-back option: Customer decisions, routing rules and system design. *Operations Research* 52(2) 271-292.

Armony, M., Shimkin, N. and W. Whitt. 2009. The Impact of delay announcements in many-server queues with abandonment. *Operations Research* 57(1) 66-81.

Burnetas, A. and A. Economou. 2007. Equilibrium customer strategies in a single server Markovian queue with setup times. *Queueing Systems: Theory and Applications* 56(3-4) 213-228.

Chen, H. and M. Frank. 2004. Monopoly pricing when customers queue. *IIE Transactions* 36(6) 1-13.

Cui, S. and S. Veeraraghavan. 2016. Blind queues: The impact of consumer beliefs on revenues and congestion. *Management Science* 62(12) 3656-3672.

Debo, L., Parlur, C. and U. Rajan. 2012. Signaling quality via queues. *Management Science* 58(5) 876-891.

Debo, L., Rajan, U. and S. Veeraraghavan. 2020. Signaling quality via long lines and uninformative prices. *Manufacturing & Service Operations Management* 22(3) 513-527.

Debo, L. and S. Veeraraghavan. 2014. Equilibrium in queues under unknown service times and service value. *Operations Research* 62(1) 38-57.

Edelson, N. and D. Hildebrand. 1975. Congestion tolls for Poisson queueing processes. *Econometrica* 43(1) 81-92.

Gentzkow, M. and E. Kamenica. 2016. A Rothschild-Stiglitz approach to Bayesian persuasion. *American Economic Review* 106(5) 597-601.

Grossman, S. J. 1981. The informational role of warranties and private disclosure about product quality. *The Journal of Law and Economics* 24(3) 461-483.

Guo, P., Haviv, M., Luo, Z. and Y. Wang. 2020. Signaling service quality through queue disclosure. Available at *SSRN* 3687086.

Guo, P. and P. Zipkin. 2007. Analysis and comparison of queues with different levels of delay information. *Management Science* 53(6) 962-970.

Guo, P. and R. Hassin. 2011. Strategic behavior and social optimization in Markovian vacation queues. *Operations Research* 59(4) 986-997.

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. *Probability in the Engineering and Informational Sciences* 21(3) 361-380.

Hassin, R. 2016. *Rational Queueing.* Chapman and Hall/CRC. London.

Hassin, R. and M. Haviv. 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. and M. Haviv. 2003. *To queue or not to queue: Equilibrium behavior in queueing systems.* Kluwer Academic Publishers. London.

Hassin, R. and R. Roet-Green. 2017. The impact of inspection cost on equilibrium, revenue, and social welfare in a single-server queue. *Operations Research* 65(3) 804-820.

Hassin, R. and R. Roet-Green. 2018. Cascade equilibrium strategies in a two-server queueing system with inspection cost. *European Journal of Operational Research* 267(3) 1014-1026.

Hu, M., Li, Y. and J. Wang. 2018. Efficient ignorance: Information heterogeneity in a queue. *Management Science* 64(6) 2650-2671.

Ibrahim, R., Armony, M. and A. Bassamboo. 2017. Does the past predict the future? The case of delay announcements in service systems. *Management Science* 63(6) 1762-1780.

Ibrahim, R. 2018. Sharing delay information in service systems: A literature survey. *Queueing Systems* 89(1-2) 49-79.

Kamenica, E. and M. Gentzkow. 2011. Bayesian persuasion. *American Economic Review* 101(6) 2590-2615.

Koutsoupias, E. and C. Papadimitriou. 1999. Worst-case equilibria. *Annual Symposium on Theoretical Aspects of Computer Science*. Springer. Berlin.

Kremer, M. and L. Debo. 2016. Inferring quality from wait time. *Management Science* 62(10) 3023-3038.

Li, K., Cui S. and J. Wang. 2020. On the optimal disclosure of queue length information. *Naval Research Logistics* forthcoming.

Lingenbrink, D. and K. Iyer. 2019. Optimal signaling mechanisms in unobservable queues. *Operations Research* 67(5) 1397-1416.

Milgrom, P. R. 1981. Good news and bad news: Representation theorems and applications. *The Bell Journal of Economics* 380-391.

Naor, P. 1969. The regulation of queue size by levying tolls. *Econometrica* 37(1) 15-24.

Rayo, L. and I. Segal. 2010. Optimal information disclosure. *Journal of Political Economy* 118(5) 949-987.

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

Veeraraghavan, S. and L. Debo. 2011. Herding in queues with waiting costs: Rationality and regret. *Manufacturing & Service Operations Management* 13(3) 329-346.

Whitt, W. 1999. Improving service by informing customers about anticipated delays. *Management Science* 45(2) 192-207.

Yu, M., Debo, L. and R. Kapuscinski. 2016. Strategic waiting for consumer-generated quality information: Dynamic pricing of new experience goods. *Management Science* 62(2) 410-435.

Yu, Q., Allon, G. and A. Bassamboo. 2017. How do delay announcements shape customer behavior? An empirical study. *Management Science* 63(1) 1-20.

Yu, Q., Allon, G. and A. Bassamboo. 2021. The reference effect of delay announcements: A field experiment. *Management Science* Articles in Advance 1-19.

Yu, Q., Allon, G., Bassamboo, A. and S. Iravani. 2018. Managing customer expectations and priorities in service systems. *Management Science* 64(8) 3942-3970.

Appendix: Proofs

Proof of Proposition 1. If a queue disclosure strategy (f_h, f_l) can induce the Bayes-plausible distribution of posteriors as presented in Proposition 1, it should satisfy that $p^R = \delta_0 f_h + (1-\delta_0) f_l$ and $p_{H|R} = P_{H|R}(f_h, f_l) = \frac{\delta_0 f_h}{\delta_0 f_h + (1-\delta_0) f_l}$. From these two equations, we obtain that $f_h = p^R p_{H|R}/\delta_0$ and $f_l = p^R(1-p_{H|R})/(1-\delta_0)$.

Proof of Lemma 1. First, $\lambda_e^R(\delta) = \lambda\left(1 - \frac{\rho^{n(\delta)}}{\sum_{i=0}^{n(\delta)} \rho^i}\right) = \mu - \frac{\mu}{\sum_{k=0}^{n(\delta)} \rho^k}$, where $n(\delta) = \lfloor [\delta V_h + (1-\delta)V_l)]\mu/\theta \rfloor$. So, $\lambda_e^R(\delta)$ increases in $n(\delta)$. Second, as δ increases from 0 to 1, $n(\delta)$ repeats the following pattern: it first remains unchanged for a while, then increases by 1, and then remains unchanged, etc. The change of $\lambda_e^R(\delta)$ in δ is a consequence of this change pattern of $n(\delta)$ in δ.

Note that when $\lambda \geq \mu - \theta/V_h$, $\lambda_e^C(\delta) = \mu - \frac{\theta}{\delta V_h + (1-\delta)V_l}$ for all $0 \leq \delta \leq 1$, which is concave and increasing in δ. When $\lambda < \mu - \theta/V_h$, $\lambda_e^C(\delta)$ consists of two pieces, first an increasing and concave function $\mu - \frac{\theta}{\delta V_h + (1-\delta)V_l}$ on the domain $\delta \in \left[0, \frac{\theta - (\mu-\lambda)V_l}{(\mu-\lambda)(V_h - V_l)}\right]$ and then a constant λ on the domain $\delta \in \left[\frac{\theta-(\mu-\lambda)V_l}{(\mu-\lambda)(V_h-V_l)}, 1\right]$. The overall function is still concave.

Proof of Proposition 2. As shown in Section 3, any queue disclosure strategy (f_h, f_l) yields a Bayes-plausible distribution of posteriors (i.e., $\delta_0 = [\delta_0 f_h + (1-\delta_0) f_l] P_{H|R}(f_h, f_l) + [\delta_0(1-f_h) + (1-\delta_0)(1-f_l)] P_{H|C}(f_h, f_l)$), and an objective value $\lambda_e(f_h, f_l) = [\delta_0 f_h + (1-\delta_0) f_l]\lambda_e^R(f_h, f_l) + [\delta_0(1-f_h) + (1-\delta_0)(1-f_l)]\lambda_e^C(f_h, f_l)$. It is straightforward to show that the point $(\delta_0, \lambda_e(f_h, f_l))$ can be regarded as the convex combination of two points $(P_{H|R}(f_h, f_l), \lambda_e^R(P_{H|R}(f_h, f_l)))$ and $(P_{H|C}(f_h, f_l), \lambda_e^C(P_{H|C}(f_h, f_l)))$, which implies that $(\delta_0, \lambda_e(f_h, f_l)) \in co(\lambda_e^R, \lambda_e^C)$. On the other hand, given $(\delta_0, \Lambda) \in co(\lambda_e^R, \lambda_e^C)$, there exist δ_1, δ_2 and $\hat{\alpha}$ such that $\hat{\alpha}\delta_1 + (1-\hat{\alpha})d_2 = \delta_0$ and $\hat{\alpha}\lambda_e^R(\delta_1) + (1-\hat{\alpha})\lambda_e^C(\delta_2) = \Lambda$ $(0 \leq \delta_1, \delta_2, \hat{\alpha} \leq 1)$. This indicates that when the prior probability for the service quality to be high is δ_0, we can always identify a queue disclosure strategy that yields the corresponding effective arrival rate Λ for the server according to Proposition 1.

Proof of Proposition 3. According to Proposition 2, under the given prior δ_0, all effective arrival rates that can be induced by feasible queue disclosure strategies constitute the set $\{\Lambda|(\delta_0, \Lambda) \in co(\lambda_e^R(\cdot), \lambda_e^C(\cdot))\}$. The conclusion then follows by the definition of $\Lambda_e(\delta_0)$.

Proof of Lemma 2. Since $\lambda_e^C(\delta) > \lambda_e^R(\delta)$ for all $\delta \in [0,1]$ and $\lambda_e^C(\delta)$ is concave in δ, all convex combinations between a point on $\lambda_e^R(\cdot)$ and a point on $\lambda_e^C(\cdot)$ fall on or below the function curve $\lambda_e^C(\cdot)$. Therefore, $\Lambda_e(\delta) = \lambda_e^C(\delta)$ and $(f_h^e, f_l^e) = (0,0)$. Similarly, when $\lambda_e^R(\delta)$ is a horizontal line and $\lambda_e^R(\delta) \geq \lambda_e^C(\delta)$ for all $\delta \in [0,1]$, these convex combinations fall on or below the flat line $\lambda_e^R(\cdot)$. Therefore, $\Lambda_e(\delta) = \lambda_e^R(\delta)$, and $(f_h^e, f_l^e) = (1,1)$.

Proof of Proposition 4. Part (i) clearly holds because, in this case, the effective arrival rate equals the potential arrival rate under the "always concealing" strategy.

For part (ii), based on the relationship between f_h and f_l, we consider two cases: $f_h < f_l$ and $f_h \geq f_l$. In the first case, $P_{H|R} < \delta_0$ (and thus, $P_{H|C} > \delta_0$). It then follows that $\lambda_e^R(0,1) \leq \lambda_e^R(f_h, f_l) \leq \lambda_e^R(1,1)$ and $\lambda_e^C(0,0) \leq \lambda_e^C(f_h, f_l) \leq \lambda_e^C(0,1)$. We know that the function $\lambda_e^C(0,1)$ becomes flat when λ increases to a certain value while $\lambda_e^R(1,1)$ always strictly increases with λ. Hence, $\lambda_e^R(1,1)$ crosses $\lambda_e^C(0,1)$ from below and exactly once as λ increases. Denote this crossing point by $\bar{\lambda}_1^e$. It then follows that as long as $\lambda > \bar{\lambda}_1^e$, $\lambda_e^R(1,1) > \lambda_e^C(f_h, f_l)$. Together with $\lambda_e^R(1,1) \geq \lambda_e^R(f_h, f_l)$, we can conclude that "always revealing" is the optimal choice for the service provider.

We now show the case where $f_h \geq f_l$. In this case, $P_{H|R} \geq \delta_0$ (and hence, $P_{H|C} \leq \delta_0$), and thus, $\lambda_e^R(1,1) \leq \lambda_e^R(f_h, f_l) \leq \lambda_e^R(1,0)$ and $\lambda_e^C(1,0) \leq \lambda_e^C(f_h, f_l) \leq \lambda_e^C(0,0)$. To show that "always revealing" is the best strategy, we first introduce an effective arrival rate function which is always no less than $\lambda_e(f_h, f_l)$, and then show that the effective arrival rate under "always revealing" can still outperform this arrival rate. Define

$$\bar{\lambda}_e(f_h, f_l) := [\delta_0 f_h + (1-\delta_0) f_l] \lambda_e^R(f_h, f_l) + [\delta_0(1-f_h) + (1-\delta_0)(1-f_l)] \lambda_e^C(0,0).$$

This new function replaces the term $\lambda_e^C(f_h, f_l)$ in the expression of $\lambda_e(f_h, f_l)$ with a larger value term $\lambda_e^C(0,0)$ and thus $\lambda_e(f_h, f_l) \leq \bar{\lambda}_e(f_h, f_l)$. Denote the optimal solution of maximizing $\bar{\lambda}_e(f_h, f_l)$ by $(\bar{f}_h^e, \bar{f}_l^e)$. Then, we have that $\lambda_e(f_h^e, f_l^e) \leq \bar{\lambda}_e(\bar{f}_h^e, \bar{f}_l^e)$.

We now show that there exists a threshold $\bar{\lambda}_2^e$ such that when $\lambda > \max\{\bar{\lambda}_1^e, \bar{\lambda}_2^e\}$, "always revealing" yields an effective arrival rate no less than $\bar{\lambda}_e(f_h^e, f_l^e)$. Recall that when the server reveals the queue length, customers join if and only if the queue length upon arrival (including themselves) is no greater than $n_e(f_h, f_l)$, where $n_e(f_h, f_l) = \lfloor V_R(f_h, f_l)\mu/\theta \rfloor$. Define a set of integers

$$S := \{n \in N_+ : n_e(1,1) < n \leq n_e(1,0)\},$$

where N_+ is the set of all nonnegative integers. Clearly, when queue is observable, customers' joining threshold $n_e(f_h, f_l)$ always falls into the set $S \cup \{n_e(1,1)\}$, because the strategy $(1,1)$ yields the lowest expected service value and the strategy $(1,0)$ yields the largest expected service value for incoming customers. Therefore, we must have $n_e(\bar{f}_h^e, \bar{f}_l^e) \in S \cup \{n_e(1,1)\}$. If the set S is empty, let $\bar{\lambda}_2^e = 0$; otherwise, we define $\bar{\lambda}_2^e$ through the following procedure. Let $n_1, \cdots, n_{|S|}$ be all the elements in set S, where $|S|$ is the cardinality of S. We now fix the joining threshold $n_e(f_h, f_l) = n_i$ $(i = 1, \cdots, |S|)$ and consider the range $\lambda > \bar{\lambda}_1^e$. Consider the constrained maximization problem as follows:

$$(\bar{f}_h^i, \bar{f}_l^i) = \arg\max_{(f_h, f_l)} \{\bar{\lambda}_e(f_h, f_l) | n_e(f_h, f_l) = n_i, \lambda > \bar{\lambda}_1^e\}.$$

According to the definition of $\bar{\lambda}_1^e$, we have that when $\lambda > \bar{\lambda}_1^e$, $\lambda_e^R(1,1) > \lambda_e^C(0,1) > \lambda_e^C(0,0)$. Also, as $f_h \geq f_l$, we have $\lambda_e^R(f_h, f_l) \geq \lambda_e^R(1,1)$. Considering these two inequalities together, we get $\lambda_e^R(f_h, f_l) > \lambda_e^C(0,0)$. With this inequality, we can then check the expression of $\bar{\lambda}_e(f_h, f_l)$. Now, the value of the term $\lambda_e^R(f_h, f_l)$ is fixed due to a fixed joining threshold n_i and the term $\lambda_e^C(0,0)$ reaching a fixed value when $\lambda > \bar{\lambda}_1^e$. Maximizing $\bar{\lambda}_e(f_h, f_l)$ then requires to maximize the term $\delta_0 f_h + (1-\delta_0) f_l$, which yields $\bar{f}_h^i = 1$ and $0 \leq \bar{f}_l^i < 1$ (note that $\bar{f}_l^i$ cannot equal 1 under the constraint $n_e(f_h, f_l) = n_i$). Therefore, within the range $\lambda > \bar{\lambda}_1^e$, $(\bar{f}_h^e, \bar{f}_l^e)$ must be $(1,1)$ or one of $(1, \bar{f}_l^i)$ $(i = 1, \cdots, |S|)$. Furthermore, we have that

$$\lim_{\lambda \to +\infty} [\lambda_e(1,1) - \bar{\lambda}_e(1, \bar{f}_l^i)] = \mu - \left\{ [\delta_0 + (1-\delta_0)\bar{f}_l^i]\mu + (1-\delta_0)(1-\bar{f}_l^i)\left(\mu - \frac{\theta}{V_l}\right)\right\} > 0.$$

Then, for $\bar{\lambda}_e(\bar{f}_h^i, \bar{f}_l^i)$, we can find a threshold for the potential arrival rate, $\bar{\lambda}_2^i$ $(\bar{\lambda}_2^i \geq 0)$, such that when $\lambda > \bar{\lambda}_2^i$, $\lambda_e(1,1) > \bar{\lambda}_e(\bar{f}_h^i, \bar{f}_l^i)$. Let $\bar{\lambda}_2^e$ be $\max\{\bar{\lambda}_2^1, \cdots, \bar{\lambda}_2^{|S|}\}$ when the set S is nonempty. It follows that when $\lambda > \max\{\bar{\lambda}_1^e, \bar{\lambda}_2^e\}$, $\lambda_e(1,1) \geq \bar{\lambda}_e(f_h, f_l) \geq \lambda_e(f_h, f_l)$ for $f_h \geq f_l$.

Finally, let $\bar{\lambda}^e := \max\{\bar{\lambda}_1^e, \bar{\lambda}_2^e\}$. We can then conclude that $\arg\max_{(f_h, f_l)} \lambda_e(f_h, f_l) = (1,1)$ for $\lambda > \bar{\lambda}^e$.

Proof of Lemma 3.

(i) Recall that $u_e^R(\delta) = \lambda \sum_{j=0}^{n_e(\delta)-1} p_j^{n_e(\delta)} \left[\delta(V_h - V_l) + V_l - \frac{(j+1)\theta}{\mu}\right]$, with $n_e(\delta) = \lfloor [\delta V_h + (1-\delta)V_l)]\mu/\theta \rfloor$. As δ increases from 0 to 1, $n_e(\delta)$ repeats the following pattern: it first remains unchanged for a while, then increases by 1, and then remains unchanged, etc. When $n_e(\delta)$ remains unchanged, $u_e^R(\delta)$ is a linear function in δ with the slope being $\lambda(V_h - V_l)\left(1 - p_{n_e(\delta)}^{n_e(\delta)}\right)$. And when $n_e(\delta)$ increases by 1 at some $\delta = \hat{\delta}$, we have that $\hat{\delta}(V_h - V_l) + V_l - \frac{n_e(\hat{\delta})\theta}{\mu} = 0$. Notice that

$p_j^{n_e(\hat{\delta})-1} < p_j^{n_e(\hat{\delta})}$ for $j = 0, \cdots, n_e(\hat{\delta}) - 2$. Then, we can get that

$$\begin{aligned}
&\lim_{\delta \to \hat{\delta}^-} u_e^R(\delta) = \lambda \sum_{j=0}^{n_e(\hat{\delta})-2} p_j^{n_e(\hat{\delta})-1} \left[\delta(V_h - V_l) + V_l - \frac{(j+1)\theta}{\mu}\right] \\
>&\lambda \sum_{j=0}^{n_e(\hat{\delta})-2} p_j^{n_e(\hat{\delta})} \left[\delta(V_h - V_l) + V_l - \frac{(j+1)\theta}{\mu}\right] = \lambda \sum_{j=0}^{n_e(\hat{\delta})-1} p_j^{n_e(\hat{\delta})} \left[\delta(V_h - V_l) + V_l - \frac{(j+1)\theta}{\mu}\right] \\
=&u_e^R(\hat{\delta}),
\end{aligned}$$

which means that $u_e^R(\delta)$ jumps down at $\delta = \hat{\delta}$.

(ii) Recall that $u_e^C(\delta) = \lambda p_e(\delta) \left[\delta(V_h - V_l) + V_l - \frac{\theta}{\mu - \lambda p_e(\delta)}\right]$ with $p_e(\delta) = 1$ if $\lambda < \mu - \frac{\theta}{\delta(V_h - V_l) + V_l}$ and $p_e(\delta) = \frac{\mu - \theta/[\delta(V_h - V_l) + V_l]}{\lambda}$ otherwise. If $\lambda \geq \mu - \frac{\theta}{V_h}$, then $p_e(\delta) = \frac{\mu - \theta/[\delta(V_h - V_l) + V_l]}{\lambda}$ for all $\delta \in [0, 1]$, which makes $u_e^C(\delta)$ constant as 0; if $\lambda \leq \mu - \frac{\theta}{V_l}$, $p_e(\delta) = 1$ for all $\delta \in [0, 1]$, and thus $u_e^C(\delta)$ is linear and increasing in δ with the slope $\lambda(V_h - V_l)$; otherwise, $p_e(\delta) = \frac{\mu - \theta/[\delta(V_h - V_l) + V_l]}{\lambda}$ for $\delta \in \left[0, \frac{\theta/(\mu - \lambda) - V_l}{V_h - V_l}\right]$, which makes $u_e^C(\delta)$ equal to 0, and $p_e(\delta) = 1$ for $\delta \in \left(\frac{\theta/(\mu - \lambda) - V_l}{V_h - V_l}, 1\right]$, which makes $u_e^C(\delta)$ linear and increasing in δ.

(iii) Now, let us compare $u_e^R(\delta)$ and $u_e^C(\delta)$ under a given δ $(0 \leq \delta \leq 1)$. First, when $\lambda \geq \mu - \frac{\theta}{\delta V_h + (1-\delta)V_l}$, we have $u_e^R(\delta) > 0$ but $u_e^C(\delta) = 0$, which directly yield $u_e^R(\delta) > u_e^C(\delta)$. Then, consider $0 < \lambda < \mu - \frac{\theta}{\delta V_h + (1-\delta)V_l}$. Under this case, we have $0 < \rho < 1$ and $p_e(\delta) = 1$. Note that in an $M/M/1/n_e(\delta)$ queue, the expected queue length is $E[L] := \sum_{j=0}^{n_e(\delta)} j p_j^{n_e(\delta)} = \frac{\rho}{1-\rho} - \frac{[n_e(\delta)+1]\rho^{n_e(\delta)+1}}{1-\rho^{n_e(\delta)+1}}$. Then, we can express $u_e^R(\delta)$ as

$$u_e^R(\delta) = \lambda[\delta V_h + (1-\delta)V_l] - \frac{(E[L]+1)\lambda\theta}{\mu} - \lambda p_{n_e(\delta)}^{n_e(\delta)} \left[\delta V_h + (1-\delta)V_l - \frac{(n_e(\delta)+1)\theta}{\mu}\right].$$

According to the definition of $n_e(\delta)$, we have $\delta V_h + (1-\delta)V_l - \frac{(n_e(\delta)+1)\theta}{\mu} < 0$. To sum up, we can get that

$$\begin{aligned}
u_e^R(\delta) - u_e^C(\delta) >& \lambda[\delta V_h + (1-\delta)V_l] - \frac{(E[L]+1)\lambda\theta}{\mu} - \lambda \left[\delta V_h + (1-\delta)V_l - \frac{\theta}{\mu - \lambda}\right] \\
=& \frac{\lambda\theta}{\mu - \lambda} - \frac{(E[L]+1)\lambda\theta}{\mu} = \frac{[n_e(\delta)+1]\rho^{n_e(\delta)+1}\lambda\theta}{(1-\rho^{n_e(\delta)+1})\mu} > 0.
\end{aligned}$$