http://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE
Vol. 66, No. 10, October 2020, pp. 4648–4666
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Efficient Inaccuracy: User-Generated Information Sharing in a Queue

Jianfu Wang,[a] Ming Hu[b]

[a] Nanyang Business School, Nanyang Technological University, Singapore 639798; [b] Rotman School of Management, University of Toronto, Toronto, Ontario M5S 3E6, Canada

Contact: wangjf@ntu.edu.sg, https://orcid.org/0000-0002-2393-8750 (JW); ming.hu@rotman.utoronto.ca, https://orcid.org/0000-0003-0900-7631 (MH)

Received: November 8, 2017
Revised: December 4, 2018; April 25, 2019
Accepted: June 7, 2019
Published Online in Articles in Advance: May 14, 2020

https://doi.org/10.1287/mnsc.2019.3447

Copyright: © 2020 INFORMS

Abstract. We study a service system that does not have the capability of monitoring and disclosing its real-time congestion level. However, the customers can observe and post their observations online, and future arrivals can take into account such user-generated information when deciding whether to go to the service facility. We perform pairwise comparisons of the shared, full, and no queue-length information structures in terms of social welfare. Perhaps surprisingly, we show that the shared queue-length information may provide greater social welfare than full queue-length information when the hassle cost of the customers entering the service facility falls into some ranges, and the shared and full queue-length information structures always generate greater social welfare than no queue-length information. Therefore, the discrete disclosure of congestion through user-generated sharing can lead to as much, or even greater, social welfare as the continuous stream of real-time queue-length information disclosure and always generates greater social welfare than no queue-length information disclosure at all. These results imply that a little shared queue-length information—inaccurate and lagged—can go a long way and that it may be more socially beneficial to encourage the sharing of user-generated information among customers than to provide them with full real-time queue-length information.

History: Accepted by Terry Taylor, operations management.
Funding: M. Hu was financially supported by the Natural Sciences and Engineering Research Council of Canada [Grant RGPIN-2015-06757] and 2019 University of Toronto–Tsinghua University Joint Research Fund.
Supplemental Material: The online appendix is available at https://doi.org/10.1287/mnsc.2019.3447.

Keywords: observable queue • unobservable queue • information sharing • service operations

1. Introduction

With the advances in information technology, people are more connected than ever. Social media like Facebook and Twitter facilitate the creation and sharing of information via virtual communities and networks. Traffic-information-sharing apps such as Waze encourage drivers to share traffic-congestion information with one another, especially for roads that are not monitored by government-funded agencies. Such user-generated information sharing is different from the full queue-length information disclosure that has been intensively studied in the literature.

The literature on service operations (see, e.g., Hassin 2016, p. 59) points out that disclosing congestion information in real time helps to better match a service provider's capacity with customers' demand and thus improves social welfare. The rationale is that, with the real-time delay information, customers will be able to make informed decisions upon arrival, so they never join a long queue or balk from a short one, thus leading to greater social welfare than when a queue cannot be seen. With a different focus, Chen and Frank (2004) suggest that real-time delay information improves the system throughput when the arrival rate is high. Their explanation is that, with a high arrival rate and no delay information, customers are relatively unlikely to join the queue. The firm would then prefer to reveal the delay information to customers so that they will always join a short queue that they would not have joined if they had no delay information. These studies lend support to the proliferation of almost real-time information disclosure in the public service systems. For example, a growing number of hospitals post their emergency-department waiting times online for patients to see. Border crossings between the United States and Canada update the congestion information online in real time.

Despite those benefits, a large fraction of service providers do not release congestion information. That is probably due to the lack of applicable technology for gathering and distributing such information. However, thanks to the thriving mobile social-network apps, the congestion experienced by one customer can be reported to others through these social platforms, as a result of customers' spontaneous sharing of information. For example, people waiting in front of a pub may post the queue length on a social platform. During the lunch break at a major academic conference, a participant may post a photo of a long queue in front of a nearby food stand to fellow participants. The Transportation Security Administration (TSA) in the United States has attempted to take advantage of the latest development in technology to establish an inexpensive information crowdsourcing and dissemination mechanism by releasing a mobile app called MyTSA, which allows air travelers to share with one another how congested different security checkpoints are at an airport.

The above user-generated shared congestion information takes the form of a snapshot of the service system with a time stamp. Future customers can see the information provided by previous customers. At the moment the congestion information is posted, it is accurate. As time goes by, it becomes less precise. Nevertheless, by checking the previously shared information, future customers may still obtain some idea of the current queue length (QL) of the system. We refer to this information structure as *shared* queue-length information (QLI), which is different from the full QLI studied in the service operations literature (see, e.g., observable queue in Naor 1969). Unlike the full QLI that discloses the QL to customers in *continuous time*, the shared QLI consists of a countable number of *discrete-time* snapshots of the service system sampled when the customers arrived. The volume of information under the shared-QLI structure is *negligible* compared with that under full QLI.

There are various user-generated information-sharing mechanisms, with different timing of sharing (e.g., at arrival or departure) and information content (e.g., real-time QL or experienced waiting time). For instance, in the pub and food-stand examples mentioned above, the QLI is shared when customers first arrive at the queue. In the TSA example, customers can share the waiting time experienced at different checkpoints when they complete the security check. In this paper, as the first attempt to investigate user-generated information sharing, we focus on the QLI shared at arrivals, which may not exactly fit some of the settings mentioned above. (In Section 5.1, we discuss an extension to investigate the model with the QLI shared at departures.) Specifically, we study the equilibrium behavior of customers with user-generated and shared QLI to address the following research question: How does this shared QLI affect the social welfare compared with the full and no QLI?

We model a service facility with a single-server queue. Upon arrival, a customer, without observing the real-time QL in the service facility, makes the *enter-or-leave* decision based on the QLI available to her (i.e., shared, full, or no QLI). If the customer decides to enter the facility and incur a hassle cost, she discovers the real-time QL, on the basis of which she then makes the *join-or-balk* decision as in an observable queue (Naor 1969).

Under the full-QLI structure, customers can access the real-time QLI broadcast by the service provider. It is an observable queue with one decision epoch. Customers will use a threshold policy to decide whether to join or balk, except that the threshold here is lower than that in Naor (1969) because the service reward is reduced by the hassle cost. Under the no-QLI structure, customers have no QLI in the first decision epoch. In this case, following Edelson and Hilderbrand (1975), we assume customers use a symmetric mixed strategy to make the enter-or-leave decision, and we can derive related service-level measures accordingly.

Under the shared-QLI structure, customers are all connected through an online information-sharing platform where the QL can be shared as public information. Regardless of whether a customer joins or balks after accessing the real-time QL, she posts the QL she observed alongside her arrival time on the online platform. In the base model, we assume that every customer who enters the facility will share QLI online. This assumption is relaxed in Section 4. Under shared QLI, because of the time lag from the previous customer, who shared the QLI to the next customer observing it, the latter perceives the real-time QL as a random variable. In other words, the shared QLI provides customers with *inaccurate* QLI, which is different from the real-time QLI in its exact value under full QLI.

Upon observing the shared QLI online, customers may not enter the facility. If the shared QLI implies that the queue is long, and therefore the expected utility of entering the facility is lower than that of leaving for an outside alternative, customers will choose not to enter. Then, every time an undesirable QL is reported, arrivals to the system are effectively shut down. After a while, if no new QLI is shared, the real-time QL is expected to approach a low level, which guarantees a positive utility for those who enter the facility. Customers will become interested in entering the facility again. This *arrival-shutdown* phenomenon is unique to the shared information structure, unlike those observable and unobservable queueing models (see, e.g., Naor 1969 and Edelson and Hilderbrand 1975, respectively). We use an approach based on the renewal theory (see, e.g., Ross 2006) to derive the social welfare analytically under the shared-QLI structure and then compare it with that under full/no QLI.

Here, we summarize our main finding. The congestion information is a mixed blessing for social welfare. On the one hand, *a little* inaccurate shared QLI is significantly more advantageous than no QLI at all. With shared QLI, customers are discouraged from entering the facility when the queue is expected to be too long to generate positive utility, and they are attracted to the facility when the queue is more likely to be short enough to provide positive utility. However, with no QLI at all, customers can join the queue blindly. Therefore, customers under shared QLI make more informed decisions than those under no QLI. On the other hand, much more abundant information (i.e., full QLI) does not necessarily lead to greater social welfare and sometimes even hurts social welfare. In comparison with full QLI, shared QLI has its advantages. Under shared QLI, during the time when everyone expects the queue to be long, even if the real-time QL drops to a low level, arrivals do not know that this has happened and are still not willing to enter the facility. Then, the queues that are not short enough to provide significant utility for joining customers are filled by customers more slowly than under full QLI. This allows long queues to diminish because the inaccurate shared QLI turns away customers only when the queues are expected to be long; hence, on average, customers under shared QLI will face a less congested system than those under full QLI. That can cause social welfare to be greater under shared QLI than under full QLI.

2. Literature Review

The literature on queueing systems with rational, utility-maximizing customers dates back to Naor (1969). The author argues that, under full QLI, self-interested customers will overload the system, and the resulting joining rate deviates from the social optimal level. Hassin and Haviv (2003) and Hassin (2016) provide comprehensive surveys for various extensions to Naor (1969).

There is a large body of literature on delay announcement to customers. Whitt (1999) shows that customers are less likely to be blocked when the delay information is provided. Guo and Zipkin (2007) consider an M/M/1 queue with three information structures: no information; partial queue-length information; and full, exact waiting-time information. They find that more information may not always improve social welfare because of customers' heterogeneous sensitivity to waiting time. This finding is further strengthened by Guo and Zipkin (2009). These papers on delay announcement all implicitly assume that service providers offer truthful information. However, customers are often unable to verify the announced congestion information. Allon et al. (2011) study customers' strategic behavior under the service provider's unverifiable delay information, and they show that cheap talk may improve the service provider's profit and customers' expected utility. More recently, Hu et al. (2018) study an M/M/1 queue where only a fraction of customers are informed of the real-time QL. The authors find that the system's throughput and social welfare may be unimodal in the fraction of informed customers. Cui et al. (2017) show that when customers are unaware of the information-disclosure policy, it is socially optimal to disclose the QL to customers when the queue is sufficiently short or sufficiently long, but not disclose when the QL is moderate. Hassin and Koshman (2017) discover that a threshold-signaling mechanism combined with a careful price selection achieves the optimal revenue in the case of a linear waiting cost. An earlier version of this study appears in Hassin and Koshman (2014). Using a Bayesian persuasion framework, Lingenbrink and Iyer (2019) show a similar result for a broader class of customer waiting costs. Ibrahim (2018) provides a comprehensive survey on this subject.

The congestion information in the papers mentioned above, regardless of its form, is communicated to customers by the service provider in real time or, equivalently, provided to everyone upon their arrival. In our paper, we study a *lagged* and *user-generated* information structure that emerges as a result of customers' spontaneous information-sharing behavior, which may serve as an inexpensive information-generating and -sharing mechanism for service providers.

Other forms of information and their disclosure have been studied in the service operations literature. Hassin (2007) considers scenarios in which the service rate, service quality, or waiting conditions are random variables that are known to the server but not to the customers, and he investigates whether the service provider is motivated to reveal these parameters to customers. Veeraraghavan and Debo (2009) study customers' inferences about the service quality through observation of the queue length, which may lead to herding in queues. Cui and Veeraraghavan (2016) study a single-server queue with customers who may have arbitrarily different beliefs about the service capacity, and they show that revealing the service-rate information can benefit revenue, but may hurt individual or social welfare.

Our work is related to several recent papers about the influence of social networks and media on service systems. Allon and Zhang (2017) study the optimal service-level differentiation in a two-period model where early adopters' experiences in the first period will be reported on social networks and internalized by all customers in the second period. Yang and Debo (2019) study the referral priority program that enables existing customers on a waiting list to gain priority access if they successfully refer new customers through their social ties. Yang et al. (2019) study the effect of search-cost reduction, enabled by service-review websites, on service providers. The authors find that reducing the search cost would increase the average waiting time for high-quality service providers and might not improve customer welfare. In contrast, we study a dynamic service system with the *transient* QLI possibly shared with later arrivals, and we compare the steady-state system performance with that under alternative information structures.

User-generated content (UGC) has been extensively studied in the marketing and information system literature; see, for example, Fader and Winer (2012) for a special issue of *Marketing Science* on why people make UGC contributions and the impact of such contributions. A large collection of papers study the impact of UGC on firms' decisions, assuming users' content-generating and -sharing behavior as being exogenously given; see, for example, Kwark and Raghunathan (2018) for the impact of UGC on product design. In this paper, the user-generated queue length information is dynamically endogenized by those who enter the facility. Moreover, our paper is related to a large body of literature on incentives for information sharing in supply chains. For example, more recently, Ha et al. (2017) consider the demand information sharing in two competing supply chains, where each supply chain is composed of one retailer and one manufacturer. The authors discover that information sharing may benefit the supply chain when the manufacturer is efficient in cost reduction. In our setting, we find that QLI sharing among fellow customers benefits the social welfare.

The work most closely related to ours is Hassin and Roet-Green (2018). In their model, customers' travel time to the service queue follows an exponential distribution. They investigate whether the service provider should provide customers with the QLI on the service queue prior to traveling. They show that to maximize throughput, it is better to disclose (respectively [resp.], conceal) the QLI if the congestion is high (resp., low) and that, to maximize social welfare, it is better to release the QLI. The information structure in our model is different. The information available to a customer at the first decision epoch was posted by the previous customer some time ago, and the customer obtains accurate real-time congestion information only if she arrives at the service facility. Moreover, our focus is on comparing the system performance measures under the shared QLI with those under full and no QLI.

3. Model

Consider a service facility that is modeled as a single-server queue. Service times are distributed exponentially with mean $1/\mu$. The demand for obtaining the service follows a homogeneous Poisson process with arrival rate Λ. Both the service rate μ and arrival rate Λ are public information. We denote by $\rho = \Lambda/\mu$ the offered load of the system. The system uses a first-come-first-served service discipline. Customers have an identical service reward R and marginal waiting cost c per unit of time, which are public information as well.

The service provider does not possess a mechanism for generating and sharing real-time queue-length information, but *all* customers have free access to an online information-sharing platform where the QL posted by someone is shared as public information. In Section 4, we relax this assumption and allow only a fraction of customers to have access to such a platform.

Hassle Cost. We assume that if customers enter the service facility, they incur an exogenous hassle cost h and observe the real-time QLI. The hassle cost represents the effort customers exert to access the queue and then obtain the real-time QLI (if it is not publicly disclosed). It is not a payment transfer from customers to the service provider. Customers *cannot* join the queue without incurring this cost. This is consistent with the literature. In retail operations literature, the hassle cost has been used to capture customers' effort for traveling to the store and searching for the product; see, for example, Gao and Su (2017). Indeed, entering the service facility may take not only effort but also time. As the first attempt to model user-generated information sharing, we focus on a parsimonious model and assume that the travel time to the facility is negligible. We refer the readers to Hassin and Roet-Green (2018) for a model that captures the travel time to a service system.

As a potential demand arises at time t, the customer makes an *enter-or-leave* decision on the basis of the hassle cost h and the latest QLI ϕ that was shared on the platform δ time units ago at time $T = t - \delta$.

Second Decision: Join-or-Balk. If an "entry" decision is made, the customer enters the facility and discovers the real-time QL, ψ, at time t (due to the assumption of negligible travel time), on the basis of which she then makes the *join-or-balk* decision. At this decision epoch, the customer's utility from joining the queue is

$$U_{join} = R - c\frac{\psi + 1}{\mu}. \tag{1}$$

If the customer chooses to balk, she obtains zero utility. Let $\lfloor x \rfloor$ denote the largest integer not exceeding x. Then, there is a threshold $n \equiv \lfloor \nu \rfloor$ where $\nu \equiv R\mu/c$, such that the customer joins the queue if and only if $U_{join} \geq 0$—that is, $\psi \leq n - 1$.

Information Sharing. Regardless of the join-or-balk decision, upon entering the facility, a customer posts her observation of the QLI along with the time stamp t on the online information-sharing platform. For convenience, we let the information posted be the total number of customers in the system after the customer makes her join-or-balk decision. For example, suppose the customer sees a number of ψ customers ahead of her. If $\psi < n$ and the customer decides to join, then she posts the QLI $\phi = \psi + 1$ on the platform; if $\psi \geq n$ and the customer balks, then she posts the QLI as $\phi = \psi$. Clearly, no customers are interested in joining a queue of length n and then post the QLI $\phi = n + 1$, so ϕ is at most n; on the other hand, any rational customer will be interested in joining an empty queue, so ϕ is at least one.

In the base model, we assume that every customer who arrives at the facility will share information online. At a service completion, the customer leaves the system without posting anything. (Note that if every customer posts the QLI on the platform upon both arrival and departure, future arrivals can infer the real-time QL accurately simply by counting, and the information structure is essentially equivalent to an observable queue, which is not our interest here.) Furthermore, in Section 5.1, we discuss how to handle the situation where customers post the QLI only at departures.

As a feature of our model, the information available to customers in the first decision epoch is not a continuous stream of real-time QL. Instead, it takes the form of a countable number of snapshots of the service system, each with a time stamp. Although future arrivals can see the information posted by earlier customers, the information is no longer accurate when they arrive. As mentioned, we refer to this user-generated information structure as the *shared*-QLI structure.

In our model, the service quality is public knowledge among customers—for example, obtained through review websites. The shared QLI does not change customers' inference of the service quality. Of course, QL can signal service quality (see, e.g., Veeraraghavan and Debo 2009), which we do not capture in our model.

First Decision: Enter-or-Leave. Let $\Psi(\delta, \phi)$ denote the random variable representing the real-time QL (including the one in service, if there is one) that can be observed by a customer arriving at time t—that is, δ time units after the most recent QLI ϕ was posted at time T. Under the assumption that all customers post information online, the realization of $\Psi(\delta, \phi)$ ranges from zero to ϕ. Recall that at the second decision epoch, customers join if and only if the observed QL is less than n. If a customer enters the facility with $\Psi(\delta, \phi) = k < n$ customers, her utility is the expected utility of joining a queue of length k less the hassle cost h. If a customer enters the facility with $\Psi(\delta, \phi) = n$ customers, she will balk, and her utility is $-h$. Thus, using the total probability theorem, we have the customer's expected utility of entering the facility:

$$
\begin{aligned}
&U_{enter}(\delta, \phi)\\
&= \begin{cases} \sum_{k=0}^{\phi} P\{\Psi(\delta, \phi) = k\} \left(R - c\dfrac{k+1}{\mu}\right) - h & \phi < n, \\ \sum_{k=0}^{n-1} P\{\Psi(\delta, n) = k\} \left(R - c\dfrac{k+1}{\mu}\right) - h & \phi = n, \end{cases}\\
&= \begin{cases} R - h - c\dfrac{E[\Psi(\delta, \phi)] + 1}{\mu} & \phi < n, \\ R - h - c\dfrac{E[\Psi(\delta, n)] + 1}{\mu} - e^{-\mu\delta}\left(R - c\dfrac{n+1}{\mu}\right) & \phi = n, \end{cases}\\
&= \begin{cases} \dfrac{c}{\mu}(\omega - 1 - E[\Psi(\delta, \phi)]) & \phi < n, \\ \dfrac{c}{\mu}\left(\omega - 1 - E[\Psi(\delta, n)] - e^{-\mu\delta}(\nu - n - 1)\right) & \phi = n, \end{cases}
\end{aligned}
\tag{2}
$$

where $\omega \equiv (R - h)\mu/\mathrm{c}$. Clearly, we have $\omega \leq \nu$. From the second equation of (2), we see that the utility of entering the facility U_{enter} is the service reward R less the combined hassle cost h and expected waiting cost when a customer actually joins the queue. This should be adjusted for the case that a customer balks after entering a facility with n customers. If the customer chooses to leave at the first decision epoch, she receives zero utility. A rational customer enters the facility if and only if $U_{enter} \geq 0$.

We write $m = \lfloor \omega \rfloor$ as the threshold such that if customers know the real-time QLI at all times (the full-QLI structure), upon arrival, a customer will enter the facility and incur the hassle cost if and only if the real-time QL is no more than $m - 1$. To ensure that customers who know the queue is empty are willing to enter the facility for the hassle cost, we should have $\omega \geq 1$; otherwise, no customers will enter the facility. Further, note that when $\omega = 1$ in (2), the expected utility of entering the facility may be negative at any time after the latest QL update. For example, if a customer reports online that she is the only one in the queue (i.e., $\phi = 1 < n$) at time T, no customers will enter the facility after that. This is because no matter how much time has passed, there is always a small possibility that this customer is still there; that is, $E[\Psi(\delta, \phi)] > 0$. Then, from (2), the expected utility of entering is negative. Hence, no customers will be interested in entering the facility after the latest QL update. Thus, for the ease of exposition, we exclude the whole interval $[1, 2)$ containing the case of $\omega = 1$ in the following analysis and focus on $\omega \geq 2$.

In our base model, we consider homogeneous customers with identical service reward R, marginal waiting cost c, and hassle cost h. In Section 5.2, we discuss the model with heterogeneous customers.

Under the shared-QLI structure, there are two separate decision epochs, and, at the second one, the hassle cost has been sunk. Thus, upon arrival, customers consider the utility function (2), where the hassle cost h is a disutility. After she enters the facility, a rational customer adopts the utility function (1) on the spot to make the join-or-balk decision, where the hassle cost is forgone as a sunk cost.

We will compare the *shared*-QLI structure with (i) the *full*-QLI structure, under which customers make the enter-or-leave decision based on a continuous stream of real-time QLI; and (ii) the *no*-QLI structure, under which customers do not have the real-time QLI, and they rely on the long-run average QL distribution to make the enter-or-leave decision. The shared-, full-, and no-QLI structures offer different forms of QLI to customers at the first decision epoch. Under full QLI, customers have the real-time QL in its exact value, which is similar to the observable queue in Naor (1969). Under shared QLI, customers view the real-time QL as a random variable derived from the user-generated shared information posted by other customers some time ago. The real-time QL is overlaid with noise over the lagged, shared QLI. Under no QLI, customers do not have the real-time QLI. Once a customer enters the facility after incurring the hassle cost h, she obtains the real-time QLI and makes the join-or-balk decision under all three information structures.

3.1. Real-Time Queue Length

We now characterize the monotonicity of the real-time QL, $\Psi(\delta, \phi)$, with respect to the elapsed time δ since the latest QL update. Because there are no arrivals but only departures during (T, t), the expected real-time QL, $E[\Psi(\delta, \phi)]$, should decrease over time. This is confirmed by the following lemma.

Lemma 1. *Given the last customer's QL update ϕ, the expected real-time QL, $E[\Psi(\delta, \phi)]$, is strictly decreasing in δ (or $t = T + \delta$).*

Customers are connected through the information-sharing platform. Upon arrival, if a customer enters the facility and sees the queue, she shares the QLI ϕ with other customers. The utility of entering the facility U_{enter} is an increasing function of δ, because the expected real-time QL, $E[\Psi(\delta, \phi)]$, is decreasing in δ after the latest QLI update at time T. This is confirmed by the following lemma.

Lemma 2. *Given the last customer's QL update ϕ, the utility of entering the facility*:

i. $U_{enter}(\delta, \phi)$ *is* strictly *increasing in δ (or $t = T + \delta$).*

ii. *At time $t = T$, $U_{enter}(0, \phi) \geq 0$ if $1 \leq \phi < m$; and $U_{enter}(0, \phi) < 0$ if $m \leq \phi \leq n$.*

iii. $U_{enter}(\delta, \phi)$ *is* strictly *decreasing in ϕ.*

From Lemma 2, we see that if the latest QL update is less than m—that is, $\phi < m$—the customer who arrives at any time after the latest QL update will enter the facility—that is, $U_{enter} \geq 0$—and use the real-time QLI to make the join-or-balk decision. When the latest online QL update is $m \leq \phi \leq n$, immediately after the latest QL update, customers have no incentive to enter the facility, because an entry leads to negative utility; that is, $U_{enter} < 0$. As time goes by, customers' utility of entering the facility U_{enter} increases [see Lemma 2(i)], and when it reaches 0, customers will want to enter the facility again. Let τ_ϕ denote the length of the time it takes for the utility of entering the facility U_{enter} to increase to zero. Then, τ_ϕ is the unique solution of the following equation with variable δ,

$$U_{enter}(\delta, \phi) = 0. \tag{3}$$

After the utility of entering the facility U_{enter} increases to zero at time $T + \tau_\phi$, customers become interested in entering the facility again. Later, at time $t \geq T + \tau_\phi$, a customer will enter the facility and see, say, ψ number of customers in the queue. If $\psi < n$, then this customer joins the queue and updates the online QLI as $\phi = \psi + 1$ [this is true even when ψ is in the range of $[m, n)$—a range of QLs for which the customer would not like to enter the facility in the first place before incurring the hassle cost h, because she forgoes the hassle cost in the second decision epoch as a sunk cost that she has paid earlier in the first decision epoch]; otherwise, she balks and updates the online QLI as $\phi = n$.

The QL update ϕ may take a value greater than m. For example, when the latest update is $\phi = m$ and before the next customer enters the facility at time t, there is no service completion; the next customer will join the queue and update $\phi = m + 1$. Similar situations can happen for $m < \phi < n$. Therefore, the shared QLI ϕ for all t can range from 1 to n.

Once some customer reports a QL $m \leq \phi \leq n$, the arrival process to the facility is effectively *shut down* for a constant time period τ_ϕ; after that, customers again become interested in making the entry decision. If the latest QL update ϕ is no more than $m - 1$—that is, $1 \leq \phi \leq m - 1$—arriving customers will not stop entering the service facility. Equivalently, the facility is shut down for a time period of length zero. Consistent with the cases of $m \leq \phi \leq n$, we adopt the convention of defining the length of the arrival-shutdown period for $1 \leq \phi \leq m - 1$ as zero—that is, $\tau_\phi = 0$.

Lemma 3. *Consider $\omega \geq 2$. The arrival-shutdown period length τ_ϕ has the following properties for $m \leq \phi \leq n$:*

i. *τ_ϕ is a decreasing function of ν.*

ii. *τ_ϕ is an increasing function of h and $\lim_{h \searrow 0} \tau_\phi = 0$.*

iii. *τ_ϕ is a decreasing function of ω.*

iv. *τ_ϕ is an increasing function of ϕ.*

To the best of our knowledge, queueing systems with *arrival shutdowns* mentioned above have not been studied before. The observable queue model (see, e.g., Naor 1969) contains a feature similar to the arrival shutdown. There, once the QL rises to the joining threshold, customers stop joining the queue until the next service completion occurs. We call it endogenous *joining shutdown* in the observable queue model to differentiate it from the arrival shutdown in our model under the shared QLI and exogenous joining shutdown in a queueing model with a finite waiting room. Queueing systems with arrival shutdowns are different from those with joining shutdowns. The joining shutdown contingently depends on the real-time QL. The length of a joining shutdown period is exponentially distributed with mean $1/\mu$. The arrival shutdown counts on the expected real-time QL. Once a long queue is reported online, an arrival-shutdown period starts, and it is of a constant length for a given posted queue length.

Moreover, queueing systems with arrival shutdowns are different from those with *service vacations* that have been widely studied (see, e.g., Takagi 1991 and references therein). The literature applies the generating function approach on the embedded Markov chain at service completions to derive the mean queue length and customers' waiting time. In a queueing system with service vacations, the number of Poisson arrivals in a vacation can be arbitrarily large, whereas in a system with arrival shutdowns, the number of service completions in an arrival-shutdown period is capped by the number of customers at the beginning of that period, which makes it difficult to directly apply and extend the generating function approach.

Remark 1. In reality, customers may not always share the QLI, but only share the QLI under some circumstances. For example, when the queue is relatively long, customers have a higher tendency and longer unoccupied time to share QLI in the form of complaints about long queues. The discussion of Lemma 2 shows that only the QLI of $m \le \phi \le n$ (the queue is relatively long) concerns future customers, whereas the QLI of $\phi < m$ (the queue is short) does not affect the arrival process at the facility. Therefore, it is equivalent to consider the information-sharing norm in which only the extreme QLI $m \le \phi \le n$ is posted. Under such a norm, suppose the latest QLI is $m \le \phi \le n$. Customers who arrive in the arrival-shutdown period will leave the system due to the anticipated negative utility of making an entry. Customers who arrive after the arrival-shutdown period will enter the facility; their rationale is: (i) If no other customers arrive after the arrival-shutdown period, then the current expected QL is less than $\omega - 1$, and the entry decision offers a positive expected utility; and (ii) if some other customers arrive after the arrival-shutdown period, given the fact that there are no QL updates greater or equal to m after the arrival-shutdown period, it is certain that these customers see less than $m - 1$ customers after they enter the facility, and the current QL must be less than m. Thus, the entry option is also desirable. This explains the equivalency of the two information structures, one with every arrival at the facility sharing QLI and the other with only the extreme QLI shared. □

3.2. Service-Level Measures

In this section, we first derive, for the shared-QLI structure, the entry rate to the facility $\bar{\Lambda}^S$, the throughput Λ^S, and social welfare S^S of the system. Recall that under the shared QLI, some customers who enter the facility may find a queue of length n, so they will not join. Thus, we have $\bar{\Lambda}^S \ge \Lambda^S$.

To derive these quantities, we consider a semi-Markov process (see, e.g., section 7.6 of Ross 2006) that enters states ϕ at time points when customers update QLI on the platform. Let π_ϕ denote its stationary probability for $\phi \in \{1, 2, \ldots, n\}$. The time period between any two consecutive transitions is a *transition cycle*, which is composed of two time intervals: the *arrival-shutdown period*, from the previous QL update (at time T) to $T + \tau_\phi$, and the *arrival-open period*, from $T + \tau_\phi$ to the next QL update. Note that the first time interval in the cycle is empty for $1 \le \phi < m$. No arrivals in the arrival-shutdown period are interested in entering the facility. At the end of the arrival-shutdown period, customers become interested in entering the facility again. The next customer will enter the facility at time $T + \tau_\phi + \exp(\Lambda)$, where $\exp(\Lambda)$ denotes a random variable following the exponential distribution with parameter Λ, and discover the real-time QL ψ. Then, if the real-time QL ψ is less than n, she joins the queue and updates $\phi = \psi + 1$; otherwise, when the real-time QL is n, which can happen if the latest QL update is $\phi = n$ and there is no service completion between these two QL updates, with probability $e^{-\mu\tau_n}\rho/(\rho+1)$, this customer balks and updates $\phi = n$. Then, another transition cycle starts with the next QL update. Clearly, the length of the arrival-shutdown period is τ_ϕ, which depends on the updated QLI ϕ. The expected length of the arrival-open period is $1/\Lambda$.

At each transition, one customer enters the facility and updates the QLI as ϕ. The customer's expected utility from joining the queue is $R - \phi c/\mu$ less the hassle cost h; otherwise, with probability $\pi_n e^{-\mu\tau_n}\rho/(\rho+1)$, this customer will balk after entering the service facility and obtain zero utility after incurring the hassle cost h.

Then, by applying renewal theory (see, e.g., Ross 2006) to the semi-Markov process, customers' entry rate to the facility $\bar{\Lambda}^S$ can be computed as one divided by the expected length of a transition cycle. The throughput Λ^S is the entry rate to the facility minus the rate of balking after an entry. Lastly, social welfare S^S can be derived by the renewal theory as the expected utility of the customer who enters the facility at the end of a transition cycle divided by the expected length of a transition cycle. Note that because our model does not involve price as a transfer payment between customers and the service provider, the social welfare only includes the customer surplus.

Proposition 1 (Shared QLI). *Under the shared-QLI structure and with a hassle cost h, the entry rate to the facility $\bar{\Lambda}^S$, the throughput Λ^S, and the social welfare S^S are given by* (*OA*.7), (*OA*.8), *and* (*OA*.9) *in the online appendix.*

If no customers share any QLI online, arrivals have no QLI whatsoever in their first decision epoch. In this case, following Edelson and Hilderbrand (1975), we assume customers use a symmetric mixed strategy to make the enter-or-leave decision. In their second decision epoch, once customers have access to the real-time QLI, they make the join-or-balk decision as in an observable queue (see, e.g., Naor 1969). At this decision epoch, customers forgo the hassle cost as a sunk cost, and they will join the queue if and only if the real-time QL is less than n. We can derive related performance measures under no QLI.

Proposition 2 (No QLI). *Under the no-QLI structure, the entry rate to the facility $\bar{\Lambda}^N$, the throughput Λ^N, and social welfare S^N are*

Case	$\bar{\Lambda}^N$	Λ^N	S^N
$0 < \rho < \rho_N$	Λ	$\mu\frac{\rho(1-\rho^n)}{1-\rho^{n+1}}$	$\mu\rho(h(\rho) - h)$
$\rho \geq \rho_N$	$\mu\rho_N$	$\mu\frac{\rho_N(1-\rho_N^n)}{1-\rho_N^{n+1}}$	0

where ρ_N is the unique solution of $h = H(\rho) \equiv \frac{c}{\mu}(\nu\frac{1-\rho^n}{1-\rho^{n+1}} - \frac{1}{1-\rho} + \frac{(n+1)\rho^n}{1-\rho^{n+1}})$. We have that

i. *ρ_N decreases in the hassle cost h;*
ii. $\lim_{h\to 0} \rho_N = \infty$;
iii. $\lim_{h\to R-c/\mu} \rho_N = 0$;
iv. *$\rho_N > \rho_L$, where ρ_L is the unique solution of $L_n(\rho) = \omega - 1$ and $L_n(\rho)$ is the expected QL under entry rate $\mu\rho$.*

Proposition 2 shows that there are two critical cutoffs, ρ_L and ρ_N, in the offered load ρ that are dependent on the hassle cost h, and further, $\rho_L < \rho_N$. If $\rho > \rho_N$, customers will join with probability ρ_N/ρ so that the effective arrival rate at the facility is $\mu\rho_N$ and customers expect zero utility. If $\rho \leq \rho_N$, all customers enter the facility expecting nonnegative utility. Moreover, if $\rho \leq \rho_L$, the resulting expected QL is no more than $\omega - 1$; otherwise, if $\rho_L < \rho \leq \rho_N$, all customers enter the facility while the resulting expected QL is greater than $\omega - 1$, a level at which, if a customer knew it in the first decision epoch, she would *not* enter the facility. Such behavior does not exist in the unobservable queue, where ρ_L and ρ_N coincide. This is because in our model under no QLI, customers can choose to balk in the second decision epoch so that the cost of entering the facility with a long queue is limited by the hassle cost, which drives up the expected QL to be greater than $\omega - 1$.

On the other hand, under the full-QLI structure, customers essentially have only one decision epoch. If their decision is to enter the facility, they incur the hassle cost h. This lowers customers' service reward to $R - h$. Therefore, to analyze customers' equilibrium behavior under the full QLI in the presence of a positive hassle cost h, we only need to assume the service reward to be $R - h$, and the same result from Naor (1969) will follow.

Proposition 3 (Full QLI). *Under the full-QLI structure, the entry rate to the facility $\bar{\Lambda}^F$, the throughput Λ^F, and social welfare S^F are $\bar{\Lambda}^F = \Lambda^F = \mu\frac{\rho(1-\rho^m)}{1-\rho^{m+1}}$ and $S^F = c\rho(\frac{1-\rho^m}{1-\rho^{m+1}}\omega - (\frac{1}{1-\rho} - \frac{(m+1)\rho^m}{1-\rho^{m+1}}))$.*

3.3. Asymptotic Case

In this section, we gain understanding of the service system under the shared-, full-, and no-QLI structures by first focusing on the asymptotic case $\Lambda = \infty$. Recall from Section 3.2 that each transition cycle under the shared QLI is composed of two time intervals: the arrival-shutdown period of length τ_ϕ and the arrival-open period of expected length $1/\Lambda$. In the asymptotic case, each transition cycle has only the arrival-shutdown period, and the arrival-open period in any transition cycle is of length zero due to the infinitely high arrival rate.

Recall from Section 3.1 that the shared QLI ranges from 1 to n. This conclusion holds a fortiori for the asymptotic case, and the QL updates under the shared QLI form a semi-Markov process with the state space $\{1, 2, \ldots, n\}$. For the QL update $1 \leq \phi \leq m - 1$, the arrival-shutdown period is of length zero; then, immediately after the QL update, another customer enters the facility, joins the queue, and obtains nonnegative utility $R - h - (\phi + 1)c/\mu$. For the QL update $m \leq \phi \leq n$, right after the arrival-shutdown period, a customer entering the facility expects to see $\omega - 1$ customers and obtains zero utility. Note from Lemma 3(iii) that the length of the arrival-shutdown period τ_ϕ increases in the online QL update ϕ, for $m \leq \phi \leq n$. This means that the more customers a transition cycle starts with, the longer the arrival-shutdown period is, and the more service completions are expected to occur during the arrival-shutdown period. As a result, the next customer entering the facility sees on average $\omega - 1$ customers in the queue.

A transition cycle starts with $m+1 \le \phi \le n$ customers only when the previous transition cycle starts with $\phi - 1$ customers and there is no service completion during the arrival-shutdown period of that cycle. Recall from Lemma 3(iv) that the length of an arrival-shutdown period τ_ϕ increases in the online QL update ϕ, for $m \le \phi \le n$. Then, the larger number of customers a transition cycle starts with, the less likely the next transition cycle is to start with a lot of customers. Hence, the steady-state probability of a transition cycle starting with $\phi + 1$ customers should be less than that of a transition cycle starting with ϕ customers, for $m \le \phi < n$. When $n - m$ is sufficiently large (e.g., $n - m \ge 5$), the probability of the transition cycle starting with n customers, π_n, becomes negligible.

Similarly, a transition cycle starts with one customer only when the previous transition cycle starts with $m \le \phi \le n$ customers, and all these ϕ customers complete services during the arrival-shutdown period of length τ_ϕ, which happens with probability $\sum_{i=\phi}^{\infty} e^{-\mu\tau_\phi}(\mu\tau_\phi)^i/i!$. When m increases to infinity, the probability of all customers completing services during an arrival-shutdown period approaches zero, as does the probability of a transition cycle starting with one customer.

When both m and $n - m$ become large, the steady-state probability mass of the online QL update concentrates around state m, and both π_1 and π_n approach zero. Note that because π_n approaches zero, almost all customers who enter the facility will join the queue. Thus, we have the expected length of any transition cycle as $\sum_{i=1}^{n} \pi_i \tau_i \approx 1/\mu$. This intuition is sensible because in this limiting case one customer is expected to enter the facility and join the queue at the end of every transition cycle, and one service completion is expected to occur in every transition cycle.

Corollary 1 (Shared QLI—Asymptotic Case)**.** *In the asymptotic case $\Lambda = \infty$, under the shared-QLI structure, the entry rate to the facility $\bar{\Lambda}^S$, the throughput Λ^S, and the social welfare S^S are given by* (*OA*.12), (*OA*.13), *and* (*OA*.14) *in the online appendix. Moreover, we have*:

i. *The social welfare under the shared-QLI S^S is nonnegative—that is, $S^S \ge 0$—where the equality holds when the hassle cost is $h = R - 2c/\mu \ge 0$.*

ii. *The throughput under the shared-QLI Λ^S approaches the service rate μ when the service reward R approaches infinity—that is,* $\lim_{R\to\infty} \Lambda^S = \mu$.

Under no QLI, when the arrival rate approaches infinity, the offered load must be greater than ρ_N given in Proposition 2. In this case, only a fraction of the customers will enter the facility so that the effective entry rate at the service facility is $\mu\rho_N$. Because customers have no QLI at their first decision epoch, they expect zero utility when entering the facility. The following corollary formalizes the above intuitions.

Corollary 2 (No QLI—Asymptotic Case)**.** *In the asymptotic case $\Lambda = \infty$, under the no-QLI structure, the entry rate to the facility $\bar{\Lambda}^N$, the throughput Λ^N, and social welfare S^N are $\bar{\Lambda}^N = \mu\rho_N$, $\Lambda^N = \mu \frac{\rho_N(1-\rho_N^n)}{1-\rho_N^{n+1}}$, and $S^N = 0$.*

Under full QLI, when the arrival rate approaches infinity, once a service completion reduces the QL to $m - 1$, a customer immediately arrives and enters the service facility. The customer obtains utility $(\omega - m)c/\mu$. This event increases the QL to m. Then, after an exponential time period with mean $1/\mu$, another service completion occurs, and the same cycle repeats. The QL essentially stays at m all the time, and the server is constantly busy with customers, so the throughput under full QLI is identical to the service rate; that is, $\Lambda^F = \mu$. The following corollary provides the asymptotic throughput and social welfare under full QLI.

Corollary 3 (Full QLI—Asymptotic Case)**.** *In the asymptotic case $\Lambda = \infty$, under the full-QLI structure, the throughput Λ^F and social welfare S^F are $\Lambda^F = \mu$, and $S^F = c(\omega - m)$.*

We see from Corollary 3 that the social welfare under full QLI S^F in the asymptotic case depends only on the fraction part of ω—that is, $\omega - m$. When the hassle cost h increases, $\omega = (R - h)\mu/c$ decreases linearly. When h increases in the range of $(R - (i+1)c/\mu, R - ic/\mu]$, ω decreases in the range of $[i, i+1)$, and S^F decreases in the range of $[0, c)$ for any integer $i \ge 2$. When the hassle cost $h \searrow R - (i+1)c/\mu$, we have $\omega \nearrow i+1$ and $S^F \nearrow c$. Hence, the social welfare under full QLI S^F is a cyclic function of the hassle cost h with a cycle length c/μ.

3.4. Comparisons

We first establish a connection between the shared-, full-, and no-QLI structures. When the hassle cost is forgone as a sunk cost at the second decision epoch, customers' expected future utility is positive: If the queue is desirably short, they join it and obtain positive utility; otherwise, in the worst case, they can balk and obtain zero utility. However, customers may choose not to enter at the first decision epoch when their expected utility of entering is less than the hassle cost. Recall that the hassle cost represents the effort customers pay to obtain the real-time QLI. When the hassle cost becomes zero, the QLI is essentially free for customers under the shared- and no-QLI structures. Then, all customers under shared or no QLI will choose to enter the facility and obtain the real-time QLI, as if the system operates under the full-QLI structure, which is equivalent to Naor's model. This insight is confirmed as follows.

Proposition 4. *When the hassle cost is zero—that is, $h = 0$—the throughput and social welfare under the shared- and no-QLI structures are identical to those under the full-QLI structure—that is, $\Lambda^F = \Lambda\frac{1-\rho^n}{1-\rho^{n+1}}$—and $S^F = c\rho(\frac{1-\rho^n}{1-\rho^{n+1}}\nu - (\frac{1}{1-\rho} - \frac{(n+1)\rho^n}{1-\rho^{n+1}}))$.*

Next, we focus on the comparison of social welfare under the shared-, full-, and no-QLI structures. Hassin (2016) suggests that the real-time QLI improves social welfare: The congestion information helps match capacity better with customer demand, so that customers never join a long queue or balk from a short one. The following theorem demonstrates that this intuition still holds in our queueing model with two decision epochs.

Theorem 1. *The social welfare under full QLI S^F is greater than that under no QLI S^N—that is, $S^F > S^N$.*

We next compare the social welfare under shared and no QLI. Under shared QLI, customers have some idea of the QL from the shared QLI posted by fellow customers. Customers are discouraged from entering the facility in the arrival-shutdown period, when the queue is expected to be too long to offer positive utility, and they are attracted to the facility in the arrival-open period, when the queue is more likely to be short enough to generate positive utility. However, customers under no QLI blindly join the queue with some probability, without any adaptation to the current state of the queue. Therefore, in this case, the social welfare under shared QLI is greater than that under no QLI. The following theorem confirms the above intuition.

Theorem 2. *The social welfare under shared QLI S^S is no less than that under no QLI S^N—that is, $S^S \geq S^N$. The equality holds only when the hassle cost is $h = 0$.*

Theorems 1 and 2 suggest that systems with the objective of increasing social welfare would strictly prefer the shared- and full-QLI structure over the no-QLI structure.

We then compare the social welfare under the shared- and full-QLI structures. We first make the comparison in an asymptotic case $\Lambda = \infty$. Recall from Section 3.3 that, in this case, each transition cycle only has the arrival-shutdown period. By focusing on the asymptotic case, we zoom in on the effect of the arrival-shutdown period on social welfare. Despite the fact that the derivation of the steady-state probability distribution π_i under shared QLI in closed form is extremely difficult, we manage to compare the social welfare under shared and full QLI without the expression of π_i. Then, building on the result obtained from the asymptotic case, we explore the comparison of the shared QLI and the full QLI for the general case $\Lambda < \infty$.

Asymptotic Case $\Lambda = \infty$ (S^S vs. S^F). Under full QLI, customers receive a continuous and accurate flow of QLI. Once a service completion occurs and reduces the QL to $m - 1$, the arrival-shutdown period ends. A customer immediately arrives and enters the service facility. This event increases the QL to m. That is, all customers who enter the facility join a queue of length $m - 1$ and obtain utility $(\omega - m)c/\mu$, whose value is small when ω is close to m.

Under the shared-QLI structure, because the latest QL update is lagged and conveys some idea about the expected QL to customers, it is not as accurate as full QLI. As a result, even if the real-time QL drops to $m - 1$ before the end of the arrival-shutdown period, no customers will discover it and enter the facility. That creates opportunity for the queue to diminish during the arrival-shutdown period before the next arrival, so that the following several customers expect positive utility.

In other words, customers overjoin the queue under the full-QLI structure, which generates negative externality for future customers. This hurts the social welfare, particularly when each joining customer obtains little utility under full QLI—that is, when $\omega - m$ is small. Thus, the social welfare under shared QLI may be greater than that under full QLI.

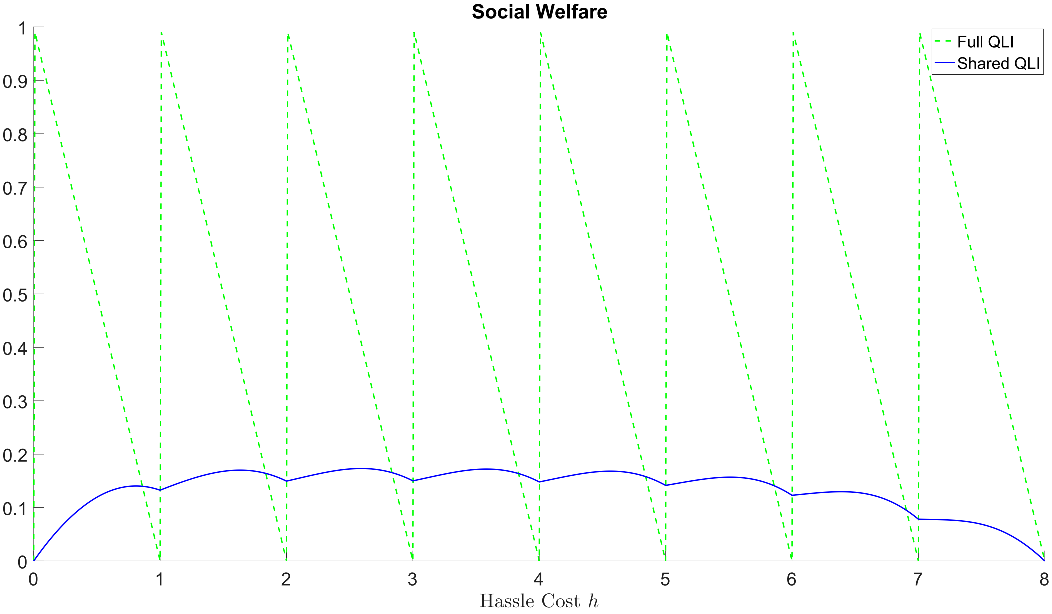
Theorem 3. *In the asymptotic case $\Lambda = \infty$, there exists an $\epsilon \in (0, c/\mu)$ such that the social welfare under shared QLI S^S is greater than that under full QLI S^F; that is, $S^S > S^F$, when the hassle cost h is in the range of $(R - ic/\mu - \epsilon, R - ic/\mu]$, for $i = 3, \ldots, n - 1$.*

Naor (1969) proposes using tolls to improve social welfare under the full-QLI structure. A toll imposed on customers who join the queue will reduce their expected net gain and make the joining shutdown period more frequent, which further alleviates congestion in the observable queue model. Theorem 3 shows that reducing the amount of congestion information available to customers may achieve the same goal by creating arrival-shutdown periods with inaccurate shared QLI. More importantly, our scheme does not involve monetary payments from customers.

To illustrate the comparison of the social welfare under shared and full QLI, in Figure 1, we plot S^S and S^F as functions of hassle cost h, in the asymptotic case $\Lambda = 10^8$. We see that, as Corollary 3 describes, the social welfare under full QLI is a cyclic function that decreases from c to zero on each band of $h \in (ic/\mu, (i + 1)c/\mu]$. Moreover, as Corollary 1 implies, the social welfare under shared QLI is positive, unless the hassle cost is very large—that is, $h = R - 2c/\mu$, or $h = 0$ for this specific example. Most importantly, on each band $h \in (ic/\mu, (i + 1)c/\mu]$, for $i = 0, 1, \ldots, 6$, there is an interval of significant length ($\approx 0.15c/\mu$) in which the social welfare under shared QLI is greater than that under full QLI.

Theorem 3 implies that, for systems with the objective of increasing social welfare, the full-QLI structure may not be more effective than the shared-QLI structure. Both the service reward R and the hassle

Figure 1. (Color online) Social Welfare as a Function of Hassle Cost h Under Full and Shared QLI with Service Reward $R = 10$, Service Rate $\mu = 1$, Marginal Waiting Cost $c = 1$, and Arrival Rate $\Lambda = 10^8$



cost h play a vital role. When the hassle cost $h \nearrow R - ic/\mu$, for any integer $i \geq 3$, the social welfare under the shared-QLI structure is greater than that under the full-QLI structure. When the hassle cost $h \searrow R - ic/\mu$, the social welfare under shared QLI may be lower than that under full QLI.

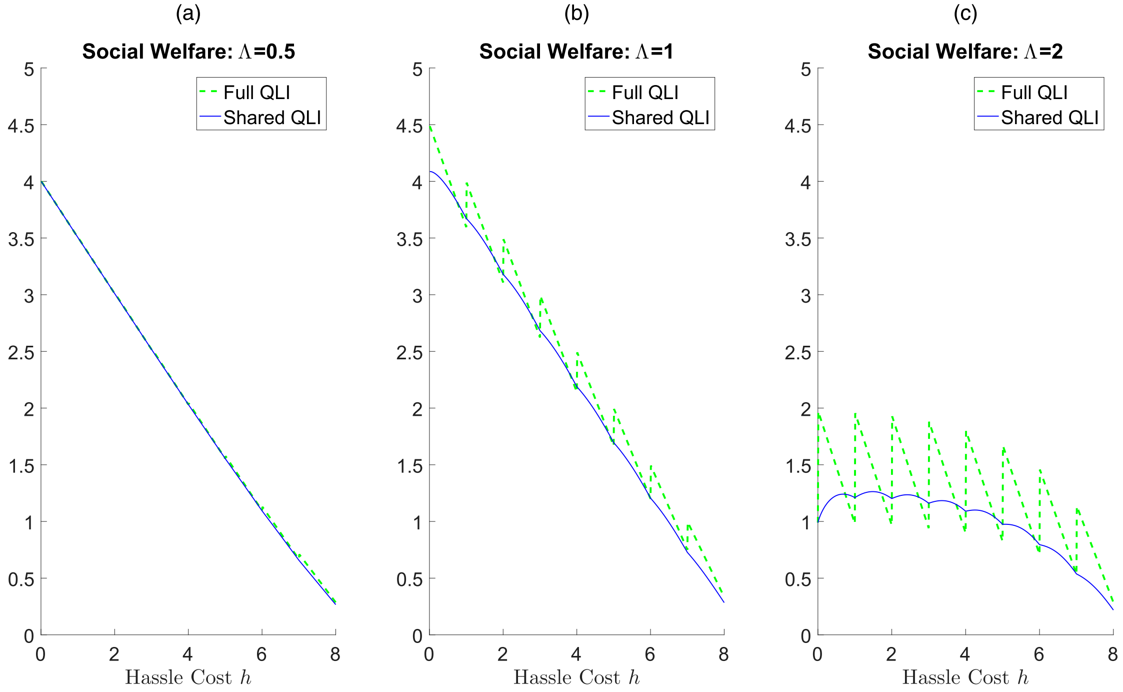
General Case (S^S vs. S^F). A finite arrival rate Λ affects the social welfare under shared- and full-QLI structures in a similar fashion to the asymptotic case. Recall from Proposition 1 that the social welfare under shared QLI can be calculated as the difference between each customer's expected service reward and waiting cost divided by the average transition-cycle length. Because the arrival rate Λ is a finite value, the arrival-open period within any transition cycle is of an expected length $1/\Lambda$. On the one hand, immediately after an arrival-shutdown period (the first interval in a transition cycle), even if customers become interested in entering the facility again, they do not arrive immediately due to a positive interarrival time. During the arrival-open period, some service completions may occur and reduce the QL seen by the next customer, whose arrival marks the end of a transition cycle. This reduces each customer's expected waiting cost and consequently increases the social welfare. We refer to this as the *wait-reduction* effect. On the other hand, when the arrival-open period within the transition cycles grows, the average cycle length increases. This effect reduces the service reward collection rate and hurts the social welfare. We refer to this as the *cycle-lengthening* effect. When the value of Λ is small, the cycle-lengthening effect is strong. For example, the expected length of the arrival-open period, $1/\Lambda$, increases from 0 to 1 if Λ decreases from ∞ to 1, whereas it increases from 1 to ∞ if Λ decreases from 1 to 0. The wait-reduction and cycle-lengthening effects also exist under the full-QLI structure.

In Figure 2, we display the social welfare under full and shared QLI, S^S and S^F, as a function of hassle cost h, for $R = 10$ in the general cases of $\Lambda = 0.5$, 1, and 2. The figure confirms that our result from the asymptotic case in Theorem 3 holds in the general case.

The wait-reduction and cycle-lengthening effects impact the social welfare under shared and full QLI similarly. For example, in Figure 2(c), when $\Lambda = 2$, for the hassle cost h in the range on the left of $R - ic/\mu$ for $i = 4, \ldots, 9$, we have $S^S > S^F$. Moreover, a low arrival rate Λ leads to long interarrival times. Then, the proportion that arrival-shutdown periods take out of the transition cycles diminishes, and the majority of customers arrive during the arrival-open period and enter the facility, as they would do under full QLI. For example, in Figure 2(a), when $\Lambda = 0.5$, the social welfare under shared QLI is almost identical to that under full QLI.

Recall that, when the arrival rate Λ decreases, the wait-reduction effect tends to boost the social welfare, whereas the cycle-lengthening effect tends to reduce it; and the interplay of the wait-reduction and cycle-lengthening effects determines the social welfare. When the arrival rate Λ is high, the cycle-lengthening effect is weak and dominated by the wait-reduction effect, so the social welfare under shared and full QLI is expected to increase when Λ decreases. This intuition is confirmed by Figure 2: The social welfare under shared and full QLI increases significantly if the arrival rate Λ decreases from 2 to 1. When

Figure 2. (Color online) Social Welfare as a Function of Hassle Cost h Under Shared and Full QLI with Service Reward $R = 10$, Service Rate $\mu = 1$, Marginal Waiting Cost $c = 1$, and Arrival Rate $\Lambda \in \{0.5, 1, 2\}$



the arrival rate Λ is low, the cycle-lengthening effect becomes strong and dominates the wait-reduction effect, so the social welfare under shared and full QLI may decrease as Λ decreases. For example, we see from Figure 2 that the social welfare under shared and full QLI decreases slightly if Λ decreases from 1 to 0.5.

Combined with Theorems 1, 2, and 3, Figure 2 offers the following managerial insight: To boost social welfare, it is more beneficial to have little, even inaccurate, information (i.e., shared QLI) than no QLI at all; however, full information does not necessarily lead to greater social welfare and may sometimes even result in less social welfare.

We also compare the throughput under these three QLI structures in Online Appendix OA1. We discover that the throughput under shared QLI is at a level similar to that under full QLI when the offered load is small and that the throughput under full QLI dominates that under shared QLI when the offered load is large. The shared-QLI structure conveys lagged and inaccurate congestion information to customers compared with the full-QLI structure. This leads to two effects on the throughput. On the one hand, the arrival-shutdown period under shared QLI allows a long queue to diminish, instead of filling up quickly after a service completion, as it would under the full-QLI structure. This input-reduction effect results in an additional possibility for the server to become idle; hence, the throughput under shared QLI may be less than that under full QLI. On the other hand, if a long queue does not diminish quickly enough in the arrival-shutdown period, customers under shared QLI may not be aware of the situation, and they may enter the facility. Then, because of the sunk hassle cost, customers may join some queues that they would have not joined under full QLI. This input boost effect leads to a higher probability of the server staying busy, so the throughput under shared QLI may be greater than that under full QLI. When the offered load is low, these two effects have similar strengths for different hassle costs, so neither the shared- nor full-QLI structure dominates the other in the throughput. However, when the offered load becomes high, the server's utilization is already close to the upper bound, and the input boost effect is curbed, so the throughput under full QLI dominates that under shared information. Last, when the offered load is high enough, the throughput under shared- and full-QLI structures strictly dominates that under no QLI.

We further compare the expected QL under shared-, full-, and no-QLI structures in Online Appendix OA2. We note that the expected number of customers at the beginning of the arrival-shutdown and arrival-open periods under shared- and full-QLI structures are expected to be close. Thus, the expected QL under shared QLI is at a level similar to that under full QLI. Furthermore, we also compare the expected QL under shared and full QLI to that under no QLI. We note that, when the offered load is in an intermediate range, higher throughput may not come at a cost of a longer waiting line: The shared- and full-QLI structures may lead to higher throughput and a lower expected QL than the no-QLI structure. A little shared information from previous customers can help to better match supply with demand, so customers enter the facility when the queue is expected to be short. When the offered load is sufficiently small or large, higher throughput is associated with a greater expected QL.

4. General Model with Shared Information

In Section 3, we assume that all customers are connected through the information-sharing platform. In this section, we extend this model by introducing a stream of *unconnected* customers, who do not have access to the shared queue-length information. The fraction of connected customers in the whole population is denoted by an exogenous parameter $\gamma \in [0,1]$, which measures the degree of social connectivity in the population. We assume that γ is common knowledge among all customers. The arrival rates (resp., offered loads) of connected and unconnected customers are $\lambda_C = \gamma\Lambda$ and $\lambda_U = (1-\gamma)\Lambda$ (resp., $\rho_C = \lambda_C/\mu$ and $\rho_U = \lambda_U/\mu$), respectively.

The connected customers behave in the same way as those under shared QLI in our model in Section 3, and they make two decisions sequentially. Upon arrival, they make the enter-or-leave decision according to the expected utility of entering the facility, which is based on the shared QLI; if they enter the facility, they make the join-or-balk decision on the basis of the real-time QLI and then post that information online.

The unconnected customers behave the same as customers under no QLI in Section 3. They have no real-time QLI before entering the service facility. In the first decision epoch, they use a symmetric mixed strategy and enter the facility with a probability $q \in (0,1)$ (resp., $q = 1$), so that the expected utility of entering the facility is zero (resp., nonnegative). Then, the unconnected customers' effective arrival rate at the facility is $q\lambda_U$. If unconnected customers choose to enter the facility at the first decision epoch, they will join the queue if the real-time QL is less than n, and otherwise balk in the second decision epoch. They do not share their observed real-time QLI.

If the degree of social connectivity is $\gamma = 1$ (resp., 0), all customers are connected (resp., unconnected), and it boils down to our base model under the shared-QLI (resp., no-QLI) structure.

4.1. Real-Time Queue Length

We now characterize the expected real-time QL given the latest QLI shared on the platform as ϕ. Between the last QL update time T and the current time $t > T$, only unconnected customers enter the facility with rate $q\lambda_U$ (otherwise, ϕ is not the latest QL update). We know that the unconnected customers are not interested in joining a queue longer than $n-1$. Hence, the QL over (T, t) should not exceed n, and the expected real-time QL at time t is determined by the transient behavior of an M/M/1 queue with arrival rate $q\lambda_U$, service rate μ, and a waiting room of size n.

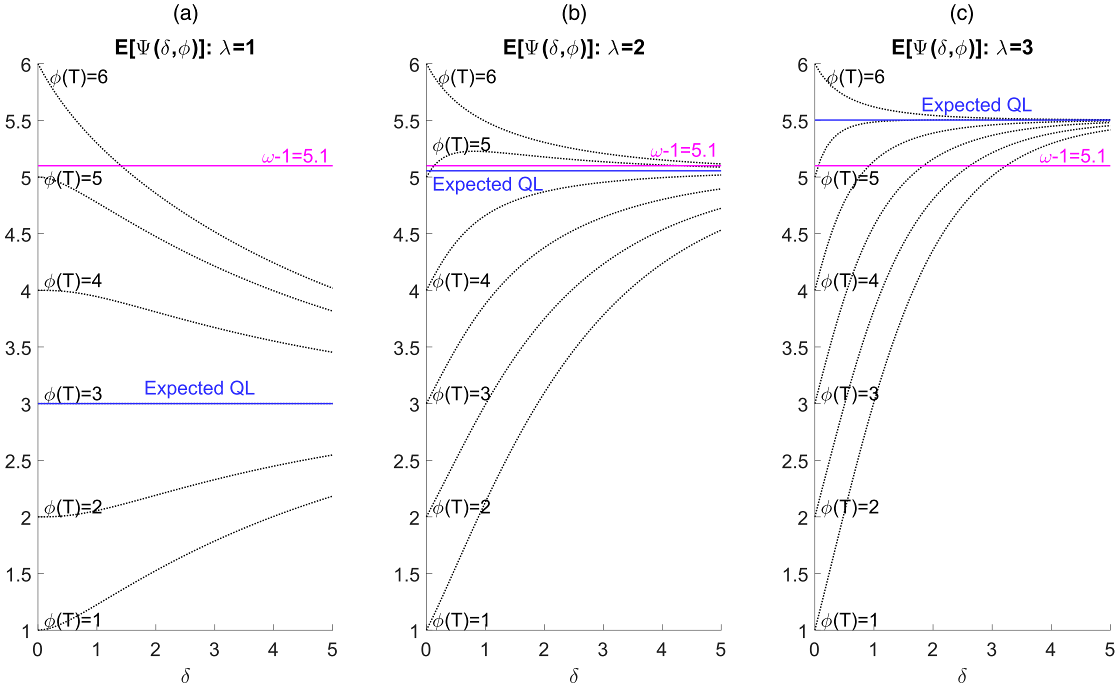
The transient behavior of an M/M/1 queue with an infinite waiting room has been discussed by Abate and Whitt (1987). To the best of our knowledge, the transient behavior of an M/M/1 queue with a finite waiting room has not been investigated in the literature. Lemma 4 in the online appendix gives a computational approach to deriving $E[\Psi(\delta,\phi)]$—the expected real-time QL at time t given the initial number of customers at time T, ϕ, in an M/M/1 queue with arrival rate λ, service rate μ, and a waiting room of size n.

Unlike in our base model, here, the expected real-time QL, $E[\Psi(\delta,\phi)]$, may not be a strictly decreasing function of t. For example, Figure 3 illustrates $E[\Psi(\delta,\phi)]$ as a function of t for $\phi \in \{1,\dots,6\}$, $\omega = 6.1$, $\mu = 1$, and $\lambda \in \{1,2,3\}$. We first make some observation from Figure 3(b). Considering the case of $\phi = 5$, we see that when t is close to T, the expected real-time QL is below $\omega - 1$, so the connected customers are willing to enter the facility; when t increases and the expected real-time QL increases from $\phi = 5$ to beyond $\omega - 1$, the connected customers arriving at t are not interested in entering the facility; as t further increases and once the expected real-time QL drops below $\omega - 1$, the connected customers will enter the facility again. For $\phi = 1,\dots,4$, the expected real-time QL is an increasing function of t, and for $\phi = 6$, the expected real-time QL is a decreasing function of t.

In all cases in Figure 3, when t approaches infinity, the expected real-time QL converges to the expected QL of an M/M/1 queue with a finite waiting room; that is, $\sum_{i=1}^{6} i(\lambda/\mu)^i / \sum_{i=0}^{6}(\lambda/\mu)^i$. In Figure 3(a), because of the low arrival rate $\lambda = 1$, the expected QL is lower than $\omega - 1$. For $\phi = 1,\dots,5$, the expected real-time QL is lower than $\omega - 1$; and for $\phi = 6$, the expected real-time QL eventually drops to below $\omega - 1$. Thus, customers who arrive long enough after the last QL update will always be interested in entering the facility. In Figure 3(c), because of the high arrival rate $\lambda = 3$, the expected QL is greater than $\omega - 1$. (Recall from Proposition 2 that customers with no QLI in the first decision epoch—that is, unconnected customers—may enter the facility at a high rate such that the resulting expected QL is greater than $\omega - 1$.) Then, for $\phi = 6$, the expected real-time QL decreases with t to the expected QL, which is always above $\omega - 1$; and it never drops below $\omega - 1$. This is to say that if any connected customer updates $\phi = 6$, no connected customers will enter the facility after that update.

Because of the complex transient behavior of an M/M/1 queue with a finite waiting room, we next resort to simulation to derive the social welfare for $0 < \gamma < 1$.

Figure 3. (Color online) Expected Real-Time QL at Time t Given that Online QLI Is $\phi(T) = 1, \ldots, 6$ Under the Setting with Parameters $\omega = 6.1$, $\mu = 1$, and $\lambda \in \{1, 2, 3\}$



4.2. Comparison

For each customer arrival rate $\Lambda \in \{0.5, 1, 2\}$, we generate 10^7 customers for an M/M/1 system. Their service time follows an exponential distribution with parameter $\mu = 1$. Customers have an identical service reward $R = 10$ and marginal waiting cost $c = 1$. The simulation ends when all 10^7 customers have their service requests satisfied or leave the system without being served because they expect the waiting time to be long.

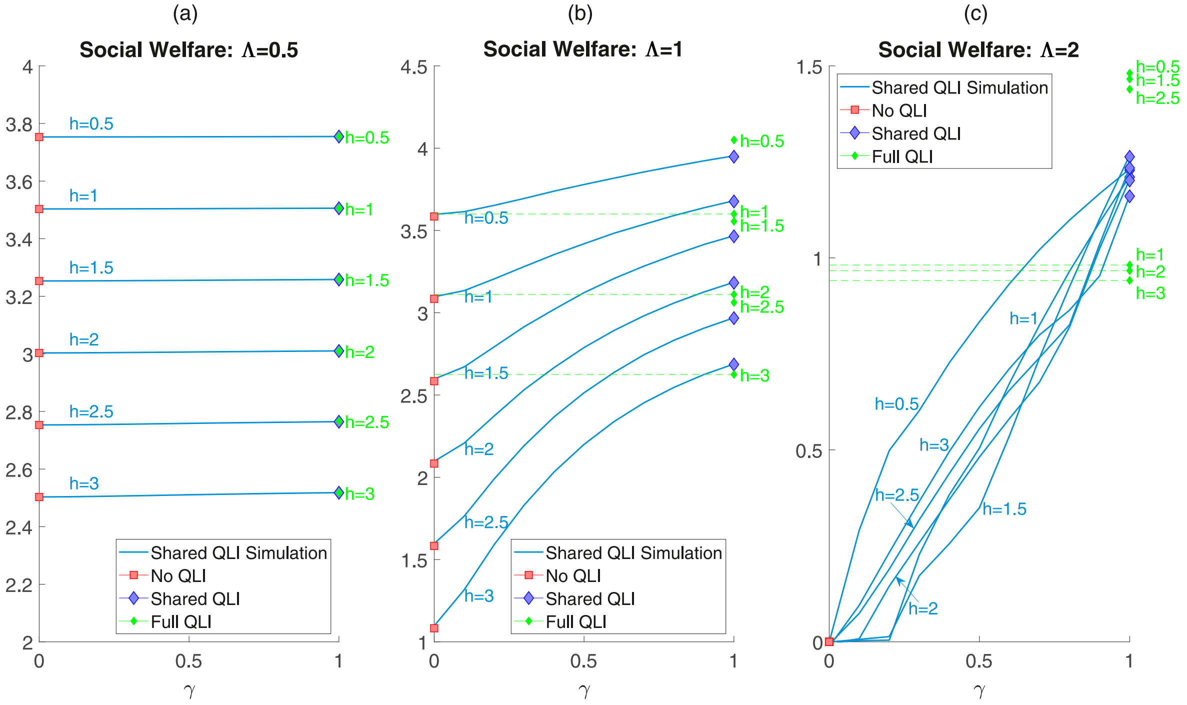
For each hassle cost h in $\{0.5, 1, \ldots, 3\}$, we randomly categorize the same group of 10^7 customers into connected and unconnected streams, according to the degree of social connectivity $\gamma \in \{0, 0.1, \ldots, 1\}$. Then, we simulate the system's social welfare with the two streams of customers. Connected customers first consider the expected real-time QL on the basis of the shared QLI in Lemma 4 upon arrival and make the enter-or-leave decision. Unconnected customers use a symmetric mixed strategy to determine the probability of entering so that the expected utility of entering the facility is nonnegative. We use a bisection search to identify unconnected customers' equilibrium entering probability. If any customer enters the facility, she will then use the real-time QL to make the join-or-balk decision.

Recall that the case $\gamma = 1$ (resp., 0) is equivalent to our base model under the shared-QLI (resp., no-QLI) structure. Thus, we can compare the social welfare of the cases $\gamma = 1$ and $\gamma = 0$ from the simulation to the theoretical ones from Propositions 1 and 2 to ensure that our simulation reaches the steady state.

Following the result of Theorem 2, we expect that when more customers share the QLI and use it to make the enter-or-leave decision, more customers will be able to avoid entering the facility with a very long queue. This improves the social welfare. Figure 4 displays the simulation result of the social welfare as a function of the degree of social connectivity along with the social welfare under shared, full, and no QLI from Propositions 1, 2, and 3. It confirms our expectation. We observe from Figure 4 that the social welfare is increasing in the degree of social connectivity γ for all cases we test. Even a small degree of social connectivity leads to an improvement beyond no QLI, and a higher degree of social connectivity results in greater social welfare.

Recall from Section 3.2 that each transition cycle contains an arrival-shutdown period and an arrival-open period in sequence. As the degree of social connectivity γ increases, the expected length of the arrival-open period $1/\lambda_C$ decreases. Moreover, connected customers will become interested in entering the service facility sooner, because there are fewer undocumented arrivals of unconnected customers. Then, the expected length of the arrival-shutdown period shortens. Hence, the expected length of transition cycles decreases with γ. Contrary to the wait-reduction and cycle-lengthening effects discussed in Section 3.4, as γ increases, we have the wait-increment effect, which reduces the social welfare, and the cycle-shortening effect, which increases the social welfare. Figure 4 shows that, as γ increases, the cycle-shortening

Figure 4. (Color online) Social Welfare as a Function of Social Connectivity $\gamma \in \{0, 0.1, \ldots, 1\}$ Under Shared, No, and Full QLI with Service Reward $R = 10$, Service Rate $\mu = 1$, Marginal Waiting Cost $c = 1$, Arrival Rate $\Lambda \in \{0.5, 1, 2\}$, and Hassle Cost $h \in \{0.5, 1, \ldots, 3\}$



(a)

(b)

(c)

effect dominates the wait-increment effect, so the social welfare under shared QLI increases.

Furthermore, we observe from Figure 4 that for some hassle costs (see, e.g., $h = 1, 2, 3$ in Figures 4(b) and 4(c)), there exists an ϵ', such that the social welfare under shared QLI is greater than that under full QLI for $\gamma \in (1 - \epsilon', 1]$. This observation is anticipated from the base model results illustrated in Figures 2(b) and 2(c), where the social welfare under shared QLI (i.e., $\gamma = 1$) is greater than that under full QLI when $h = 1, 2, 3$.

In our model, the degree of social connectivity γ is an exogenous parameter, and customers do not decide whether to become connected through the information-sharing platform. If we endogenize customers' decision of becoming connected to share and obtain the latest congestion update, the degree of social connectivity γ can be viewed as an outcome of customers' symmetric strategic behavior of deciding whether to be connected on the platform. In Online Appendix OA3, we investigate how customers' strategic behavior evolves when an information-sharing platform is introduced to a service facility that originally had no QLI disclosure (i.e., $\gamma = 0$). We discover that if the total offered load is not high, the platform will gain popularity over time, and all customers will prefer to join it. Moreover, the more customers are connected, the greater the social welfare. This insight provides support for the prevalence of some information-sharing apps. For example, the app Waze has become increasingly popular among Uber drivers. Otherwise, if the offered load is high, customers may not join the platform spontaneously. Within a certain range of degrees of connectivity, unconnected customers may relentlessly enter the facility expecting nonnegative utility, which causes excessive congestion and drives away connected customers. In this case, the social planner needs to intervene to get sufficient customers connected on the platform so that the rest of the population will voluntarily follow, increasing the social welfare as an outcome.

We caution that the underlying assumption behind connected customers is that they come to an agreement whereby they are obligated to share QLI in exchange for QLI shared by others. Nevertheless, it is challenging at a practical level to enforce spontaneous sharing of information among customers. Some customers may act as information free-riders, who use information shared by others but do not voluntarily share information to the rest. An interesting future research direction will be exploring endogenous information sharing *without* contractual sharing obligations.

5. Discussions

5.1. Shared Information at Departure

As the first attempt to investigate the case of user-generated information online, we assume customers the queue-length information observed when entering the service facility. One can also investigate other forms of shared information available to customers. For example, instead of customers sharing the QLI when they enter the facility, they may share the QLI at departure (QLID). Another similar information structure is that customers share the *sojourn time* they experienced in the system at departure. The resemblance comes from the fact that at a customer's departure, all customers in the system arrived during the departing customer's sojourn time, so her experienced sojourn time and the QL at departure are stochastically linked.

We assume that all customers share the QLID. Upon arrival, a customer knows there are no service completions since the latest QL update ϕ, so the real-time QL does not decrease. If the shared QL is sufficiently long—that is, $\phi \geq m$—due to the negative expected utility of entering the facility, no customers will enter the facility after T until the next QL update. If the QL update shared online is low—that is, $\phi < m$—the customer who arrives immediately after the update will enter the facility, because the expected utility of entering the facility is positive—that is, $U_{enter} > 0$. If every customer who arrives after the last QL update enters the facility, the expected real-time QL increases, and the expected utility of entering the facility will decrease to zero over time. If U_{enter} decreases to zero before the next QL update, no customers will enter the facility before the next QL update, due to the negative expected utility of entering. Hence, only during a time interval after a low QL update—that is, $\phi < m$—will customers enter the facility. These *arrival-turned-on periods* contrast with the arrival-shutdown periods in our model with the QLI shared when customers enter the facility. However, the same technique from Section 3 can be applied to obtain the throughput and social welfare of such service systems.

From the results obtained from Sections 3.4 and 4.2, we expect the shared QLID to dominate the no-QLI structure in social welfare. This is because customers under shared QLID only enter the facility when the queue is expected to be short and avoid entering the facility when the queue is reported and perceived to be long. This behavior not only provides customers with a higher probability of obtaining positive utility, but also reduces the negative externality they may impose on future customers.

Similar to the result from Section 3.4, the social welfare under shared QLID is expected to be at a level similar to that under full QLI. In the asymptotic case $\Lambda = \infty$, the shared-QLID structure is identical to the full-QLI structure. To see this, consider a facility under the shared-QLID structure when the latest shared QLID is m. No customers will enter the facility before the next QLID update. Once a service completion occurs, the QLID $m-1$ is shared online. One customer immediately enters the facility, and this increases the real-time QL to m. Then, the same cycle repeats. Under the full-QLI structure, in a facility with the latest QL update m, the same events occur. Thus, these two information structures are identical in the asymptotic case $\Lambda = \infty$.

5.2. Heterogeneous Customers

In our base model, we consider homogeneous customers with identical service reward R, waiting cost c, and hassle cost h. In this section, we discuss how heterogeneity may affect those insights derived in our base model. Here, we examine the model with heterogeneous hassle costs as an example. (Heterogeneity in service rewards or marginal waiting costs can be examined similarly.) We assume that there are high- and low-type customers in the service system, and the hassle cost of high-type customers, h_H, is higher than that of the low-type customers, h_L. Clearly, because of low-type customers' low hassle cost, their entries to the service facility lead to higher social surplus than high-type customers' entries with everything else being equal.

Different customers react differently to the shared QLI. Recall from Lemma 3(ii) that the length of the arrival-shutdown period increases with the hassle cost. There are two implications of this monotonicity property. First, given the identical QLI shared by previous customers, because $h_H > h_L$, high-type customers' arrival-shutdown period is longer than that of low-type customers. Then, high-type customers enter the service facility less often than the low-type customers. Second, when h_H increases and h_L decreases, high-type customers whose entries to the service facility lead to lower social surplus enter less often. Then, if low-type customers whose entries to the service facility lead to higher social surplus enter the facility more often, the social welfare will increase, unless the low-type customers' entry rate to the service facility is curbed—that is, when the low-type customers' total arrival rate λ_L is relatively low and they already enter the facility often.

We use a simulation to verify our intuition mentioned above. Let $\eta \in [0, 1]$ denote the fraction of high-type customers in the whole population, so the arrival rates of high- and low-type customers are $\lambda_H = \eta\Lambda$ and $\lambda_L = (1-\eta)\Lambda$, respectively. We define high- and low-type customers' hassle costs as $h_H = h + \xi_H$ and $h_L = h - \xi_L$, respectively, where $\xi_L, \xi_H \in [0, \min(h, R - 2c/\mu - h)]$ are the deviations of high- and low-type customers' hassle cost from h. We assume $\xi_H = \xi_L(1-\eta)/\eta$, so that the average hassle cost across all customers stays at $(\lambda_H h_H + \lambda_L h_L)/\Lambda = h$. Note that when $\xi_L = 0$, we have $h_H = h_L = h$, which boils down to our base model with homogeneous customers.

Figure 5. (Color online) Social Welfare as a Function of ξ_L Under Shared QLI with Service Reward $R = 10$, Service Rate $\mu = 1$, Marginal Waiting Cost $c = 1$, Average Hassle Cost $h = 4$, and Arrival Rate $\Lambda \in \{0.5, 1, 2\}$

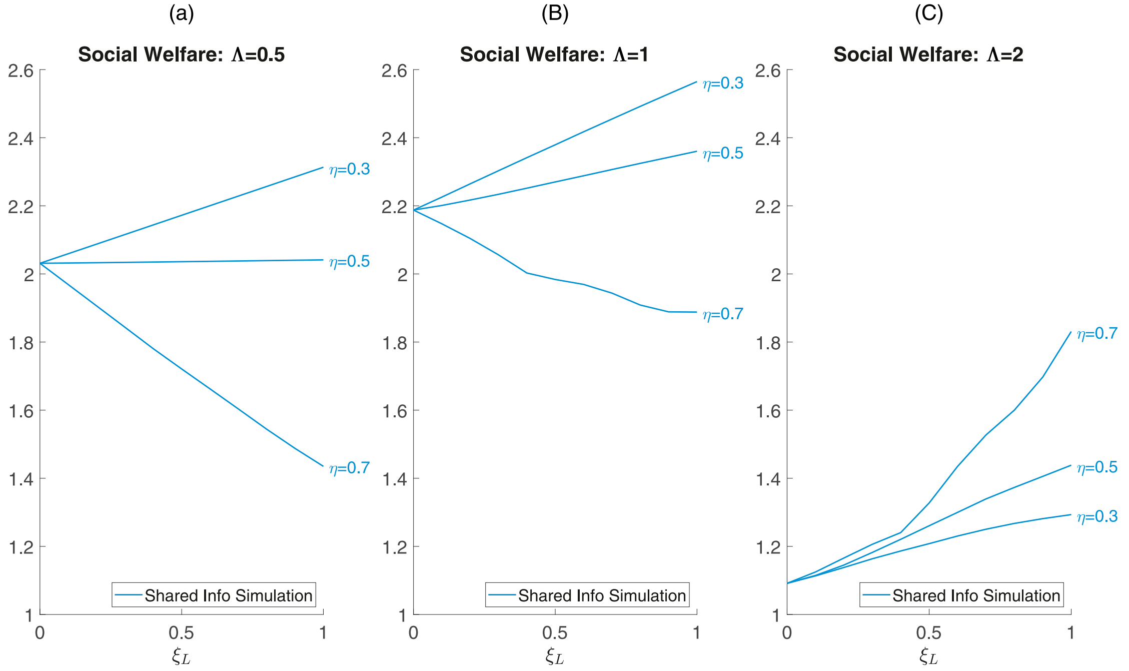


Figure 5 displays the social welfare under shared QLI as a function of ξ_L, with the service reward $R = 10$, service rate $\mu = 1$, marginal waiting cost $c = 1$, average hassle cost $h = 4$, fraction of high-type customers $\gamma \in \{0.3, 0.5, 0.7\}$, and arrival rate $\Lambda \in \{0.5, 1, 2\}$. It confirms our intuition. We observe from Figure 5 that in most cases the heterogeneity in the hassle cost improves social welfare—that is, the social welfare increases with ξ_L—except when $\eta = 0.7$ in Figures 5(a) and 5(b). In these two cases, the total arrival rate Λ is not high—that is, $\Lambda = 0.5$ and 1—so there is not much room left for the low-type customers to increase their entry rate. This confirms the robustness of the arrival-shutdown phenomenon analytically demonstrated for homogeneous customers.

Combined with comparative statics in Section 3.4, our observation here suggests that the shared QLI may still dominate full QLI in the case of heterogeneous hassle costs, at least when the heterogeneity is low enough—that is, ξ_L is sufficiently small.

5.3. Strategic Wait

Customers in our base model have two options at the first decision epoch: enter or leave. In reality, customers may have other options. For example, customers may recheck the shared information after some time, hoping to see a lower expected number of customers in the service facility. Cui et al. (2019) investigate a model of rational retrials in queues. Furthermore, customers may strategically wait outside the service facility until the expected real-time QL in the service facility is low enough, and then enter (assuming that customers form a line and follow first-come-first-served queueing discipline outside the facility). Lariviere and Van Mieghem (2004) consider a game in which customers who are delay-sensitive and try to avoid congestion choose arrival times strategically, and they show that the resulting endogenous arrival pattern approaches a discrete-time Poisson process as the number of customers and arrival points gets large. Similarly, in our model, customers' strategic waiting behavior can result from their rational decisions under the shared-QLI structure.

If the cost of strategic waiting outside the service facility is negligible, all customers under shared (resp., full) QLI will wait until the queue inside the facility is expected (resp., known) to be short. However, customers' behavior under no QLI is not affected by this waiting option introduced in the first decision epoch, because waiting does not improve customers' expected utility when there is no information. Moreover, when the arrival rate is no lower than the service rate—that is, $\Lambda \geq \mu$—the model behaves as our model in the asymptotic case $\Lambda = \infty$ in Section 3.3. This is because customers are willing to wait outside the service facility for an infinitely long time due to the zero waiting cost. When the arrival rate is less than the service rate—that is, $\Lambda < \mu$—customers' strategic waiting smooths the entry rate to the service facility. In this case, customers under the shared or full QLI have strategic waiting as a better option than leaving the service facility with zero utility, so the social welfare in this case should be greater than that in our model. From the fact that the social welfare under shared QLI is close to that under full QLI when the arrival rate is small, we anticipate a small difference between the social welfare under shared and full QLI in the model with strategic waiting at the first decision epoch.

If the cost of strategic waiting outside the service facility is positive, customers under shared or full QLI may follow a threshold policy to decide whether to wait at each time they recheck the information. To analytically track the impact of strategic waiting on our model with shared QLI, one may need to introduce another dimension in the Markov chain to record the number of customers who are waiting outside the service facility for the right moment to enter.

6. Conclusion

The proliferation of information sharing enabled by technological innovation creates an opportunity for service providers to generate and disseminate congestion information in a cost-efficient way by taking advantage of user-generated information and its sharing. In this paper, we study a single-server facility where customers share the congestion information with one another in the form of a snapshot of the system. Hence, future customers, before entering the facility and observing the real-time QL with a hassle cost, can make the enter-or-leave decision by referring to shared information.

We find that, although the shared queue-length information is most likely *lagged* and *inaccurate*, it gives customers some idea of the congestion level so they can avoid entering the facility with potentially long queues. Hence, for any arrival rate, the shared-QLI structure dominates the no-QLI structure in social welfare. We verify this analytically. Moreover, the information about long queues shared online discourages arrivals, thereby allowing long queues to diminish rather than to start filling up soon after the next service completion, as likely happens under the full-QLI structure. Therefore, the shared-QLI structure may even result in less congestion and greater social welfare than the full-QLI structure.

Our results imply that for service providers that do not have the capacity to generate and disseminate the real-time QLI, investing in such capability may not be necessary. Instead, an information-sharing platform can be a cost-efficient solution, and user-generated information sharing may lead to greater social welfare than full or no QLI. Moreover, for public service providers who do have the capacity to disclose the full congestion information to all customers, it may not be optimal to do so continuously. Our results suggest that public service providers that practice proactive information control may also benefit from periodic information release.

Acknowledgments

The authors thank the department editor, Terry Taylor, the anonymous associate editor, and three reviewers for their constructive and insightful suggestions, which significantly improved the paper. The authors also thank Mor Armony, Shiliang Cui, Laurens Debo, Pengfei Guo, Refael Hassin, Moshe Haviv, Rouba Ibrahim, Krishnamurthy Iyer, Ricky Roet-Green, Senthil Veeraraghavan, and Luyi Yang for their suggestions to improve the paper.

References

Abate J, Whitt W (1987) Transient behavior of the M/M/1 queue: Starting at the origin. *Queueing Systems* 2(1):41–65.

Allon G, Zhang DJ (2017) Managing service systems in the presence of social networks. Working paper, University of Pennsylvania, Philadelphia.

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.

Chen H, Frank M (2004) Monopoly pricing when customers queue. *IIE Trans.* 36(6):569–581.

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

Cui S, Li K, Wang J (2017) On the optimal disclosure of queue length information. Working paper, Georgetown University, Washington, DC.

Cui S, Su X, Veeraraghavan S (2019) A model of rational retrials in queues. *Oper. Res.* 67(6):1699–1718.

Edelson N, Hilderbrand D (1975) Congestion tolls for Poisson queuing processes. *Econometrica* 43(1):81–92.

Fader PS, Winer RS (2012) Introduction to the special issue on the emergence and impact of user-generated content. *Marketing Sci.* 31(3):369–371.

Gao F, Su X (2017) Omnichannel retail operations with buy-online-and-pick-up-in-store. *Management Sci.* 63(8):2478–2492.

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

Guo P, Zipkin P (2009) The effects of the availability of waiting-time information on a balking queue. *Eur. J. Oper. Res.* 198(1):199–209.

Ha AY, Tian Q, Tong S (2017) Information sharing in competing supply chains with production cost reduction. *Manufacturing Service Oper. Management* 19(2):246–262.

Hassin R (2007) Information and uncertainty in a queueing system. *Probab. Engrg. Inform. Sci.* 21(03):361–380.

Hassin R (2016) *Rational Queueing* (CRC Press, Boca Raton, FL).

Hassin R, Haviv M (2003) *To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems*, International Series in Operations Research and Management Science, vol. 59 (Springer, Berlin).

Hassin R, Koshman A (2014) Optimal control of a queue with high-low delay announcements: The significance of the queue. *Proc. 8th Internat. Conf. Performance Evaluation Methodologies Tools VALUETOOLS '14* (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, Brussels), 233–240.

Hassin R, Koshman A (2017) Profit maximization in the M/M/1 queue. *Oper. Res. Lett.* 45:436–441.

Hassin R, Roet-Green R (2018) The armchair decision: On queue-length information when customers travel to a queue. Working paper, Tel Aviv University, Tel Aviv.

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 Systems* 89(1-2):49–79.

Kwark Y, Raghunathan S (2018) User-generated content and competing firms' product design. *Management Sci.* 64(10): 4608–4628.

Lariviere MA, Van Mieghem JA (2004) Strategically seeking service: How competition can generate Poisson arrivals. *Manufacturing Service Oper. Managment* 6(1):23–40.

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.

Ross SM (2006) *Introduction to Probability Models*, 9th ed. (Academic Press, Inc., Orlando, FL).

Takagi H (1991) *Queueing Analysis: A Foundation of Performance Evaluation* (North-Holland, Amsterdam).

Veeraraghavan S, Debo LG (2009) Joining longer queues: Information externalities in queue choice. *Manufacturing Service Oper. Management* 11(4):543–562.

Whitt W (1999) Improving service by informing customers about anticipated delays. *Management Sci.* 45(2):192–207.

Yang L, Debo LG (2019) Referral priority program: Leveraging social ties via operational incentives. *Management Sci.* 65(5): 2231–2248.

Yang L, Debo LG, GuptaV (2019) Search among queues under quality differentiation. *Management Sci.* 65(8):3605–3623.