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

MANAGEMENT SCIENCE
Vol. 64, No. 6, June 2018, pp. 2650–2671
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Efficient Ignorance: Information Heterogeneity in a Queue

Ming Hu,[a] **Yang Li,**[b] **Jianfu Wang**[c]

[a] Rotman School of Management, University of Toronto, Toronto, Ontario M5S 3E6, Canada; [b] CUHK Business School, Chinese University of Hong Kong, Shatin, N.T., Hong Kong; [c] Nanyang Business School, Nanyang Technological University, Singapore 639798

Contact: ming.hu@rotman.utoronto.ca (MH); liyang@cuhk.edu.hk (YL); wangjf@ntu.edu.sg (JW)

Received: July 25, 2014
Revised: August 5, 2015; May 2, 2016; November 7, 2016
Accepted: January 4, 2017
Published Online in Articles in Advance: June 13, 2017

https://doi.org/10.1287/mnsc.2017.2747

Abstract. How would the growing prevalence of real-time delay information affect a service system? We consider a single-server queueing system where customers arrive according to a Poisson process and the service time follows an exponential distribution. There are two streams of customers, one informed about real-time delay and the other uninformed. The customers' uninformed behavior may be due to information ignorance or rational behavior in the presence of an information fee. We characterize the equilibrium behavior of customers with information heterogeneity and investigate how the presence of a larger fraction of informed customers affects the system performance measures, i.e., throughput and social welfare. We show that the effects of growing information prevalence on system performance measures are determined by the equilibrium joining behavior of *uninformed* customers. Perhaps surprisingly, we find that throughput and social welfare can be *unimodal* in the fraction of informed customers. In other words, some amount of information heterogeneity in the population can lead to more efficient outcomes, in terms of the system throughput or social welfare, than information homogeneity. For example, under a very mild condition, throughput in a system with an offered load of 1 will always suffer if there are more than 58% of informed customers in the population. Moreover, it is shown that for an overloaded system with offered load sufficiently higher than 1, social welfare always reaches its maximum when some fraction of customers is uninformed of the congestion level in real time.

History: Accepted by Gad Allon, operations management.
Funding: The first author is grateful for financial support from the Natural Sciences and Engineering Research Council of Canada [Grant RGPIN-2015-06757].
Supplemental Material: The online appendix is available at https://doi.org/10.1287/mnsc.2017.2747.

Keywords: service operations • queueing games • heterogeneous delay information • social welfare • system throughput

1. Introduction

In today's service industries, information about delays is ubiquitous. The waiting time to cross the border between the United States and Canada is posted online and updated in real time. Information about traffic jams on major roads is distributed on radio, television, and the Internet. Thanks to traffic-information-sharing apps such as Waze, real-time information about traffic may also be available even for roads that are not covered by government-funded traffic detection and monitoring.

Nevertheless, regardless of how widely available information about real-time delays may be, a large percentage of customers are still uninformed, for various reasons that may be hard to ascertain. For example, not everyone has a mobile device that might make information acquisition almost effortless, or people may simply overlook up-to-the-minute information about delays before setting out. In an online poll some 20,000 participants were asked, "How do you most often check traffic information before going out?"; 47% answered, "I don't check"; the rest checked various sources, such as TV, radio, computer, and mobile device.[1] Some people may simply be overconfident that they will be lucky. Others may sometimes check for information but not always. Another reason for information heterogeneity could be that small service providers may be unable to afford the technology for tracking and reporting the congestion levels. In that case, only drop-in customers can see the queue, whereas many potential customers cannot.

Thus, it is evident that many of today's service environments are characterized by customer heterogeneous knowledge about delays: Some are informed about the real-time delay; others are not. It is essential to understand the interaction among customers with information heterogeneity to answer the question: How do system throughput and social welfare change as the real-time delay information becomes more prevalent due to advances in information technology?

On the one hand, Chen and Frank (2004) show, by comparing full real-time delay information with no such information, that delay information is a double-edged sword for system throughput. When the system load is low, customers might be turned away with real-time information, but if they are uninformed they are likely to stay. The throughput of an observable

queueing system exceeds that of the unobservable counterpart only when the system load is high. In a more common situation where some, but not all, customers are informed, would the system throughput outperform its counterparts in the two extreme information structures, i.e., full and no information?

On the other hand, again by comparing full and no real-time delay information, Hassin (1986) argues that delay information can improve social welfare. The intuition is that congestion information helps better match capacity with customer demand: Customers never join a long queue or balk from a short one. That rationale is consistent with the prevalence of real-time delay information in today's public service industries. However, as we argue, it may be unrealistic to expect that all customers have access to real-time delay information even if it is readily available. More important, does the system inevitably suffer efficiency loss from the presence of uninformed customers?

To answer those questions, we study a single-server queue as in Hassin (1986) but with a middle ground by assuming a mix of two streams of customers who are different in their information structures. Specifically, the server posts the actual queue length in real time. One stream of customers, which we call *informed*, makes the decision to join or balk on the basis of the real-time delays. The other stream, which we call *uninformed*, is unaware of the real-time information and bases the join-or-balk decision on the average delay. In the base model, we assume that the fraction of informed customers, (the *information level*), is *exogenous*, and we study comparative statics with respect to the information level, i.e., the influence of a larger informed fraction (i.e., growing information prevalence) on system performance such as throughput and social welfare.

Given the difference in the delay knowledge, the two streams of customers have different self-interested joining behavior. Informed customers use a threshold policy: If the queue is observed to be shorter than a particular threshold, they join it; otherwise, they balk. Uninformed customers, by contrast, are aware of the expected waiting time, and they randomize their decisions between joining and balking. Although informed customers use the same state-dependent threshold strategy as they would in an observable queue (see, e.g., Naor 1969), the presence of uninformed customers undoubtedly influences the likelihood that an informed customer will join the queue. This interaction is not captured by observable or unobservable models (for the latter, see, e.g., Edelson and Hilderbrand 1975). However, our results reveal that service providers who ignore this interaction between the two segments may miss the opportunity to achieve better system performance. In particular, we show that unless the customer arrival rate is extremely low relative to the speed of service, it is possible to improve throughput or social welfare when only a fraction of customers are informed of the real-time delay. Moreover, the system performances crucially depend on the equilibrium joining behavior of *uninformed* customers.

We show that the ubiquity of delay information may have a positive or negative effect on the system throughput. In particular, we prove that there are two critical levels of offered loads. If the offered load is above (or below) the higher (or lower) one, the throughput always increases (or decreases) in the information level. These results are consistent with the comparison of the full- and no-information models in Chen and Frank (2004). However, we also show that if the offered load falls in the *intermediate* range between the two critical levels, the throughput is always *unimodal* in the information level. In addition, the throughput reaches its maximum at the information level in which *all* uninformed customers are about to adopt an always-join strategy. This finding implies that treating all customers equally as informed or as uninformed may fail to achieve the potential value of effective information control.

Contrary to the conventional wisdom that real-time congestion information always improves social welfare, we further demonstrate that social welfare is *unimodal* in the information level when the system experiences a high enough offered load. (Only when the offered load is relatively low does information prevalence always benefit social welfare.) This is because growing information prevalence has positive and negative effects on social welfare. On the positive side, if real-time system congestion is visible to customers, system capacity can be more efficiently matched with customer demand intertemporally because informed customers seek service only when the queue is short enough to yield a surplus. However, informed customers' selfish joining behavior may overload the system, especially when the customer arrivals are overwhelming. In this situation, the presence of uninformed individuals in fact mitigates the system congestion: Uninformed customers are reluctant to join a busy system without real-time information. This disincentive helps free up the capacity to serve more of the informed customers, who contribute more surpluses to social welfare. Nonetheless, when a large proportion of a high customer volume is informed, uninformed individuals eventually lose interest in the service. If they all choose to balk, as information prevalence grows, the system suffers from rising externality inflicted by an increasing fraction of informed individuals. Hence social welfare deteriorates as a result of growing information prevalence.

Our results highlight the fact that some degree of real-time information heterogeneity in the population can lead to more *efficient* outcomes in terms of system throughput or social welfare than can information homogeneity. The presence of uninformed customers does not necessarily harm the system. In fact,

it improves system throughput when the system experiences low offered loads and increases social welfare when the system experiences high offered loads.

Our results also imply that there may be value in intentionally introducing information heterogeneity and controlling the availability of real-time delay information. For tractability, our base model studies comparative statics by assuming an exogenous fraction of uninformed customers. In Section 5, we propose two strategies for achieving the optimal performance. For the case where service providers are in charge of information disclosure, we suggest offering different granularities, real-time or expected, of delay information. When customers are fully rational and make self-interested joining decisions, we discuss how to achieve a socially optimal degree of information heterogeneity through an information access fee. In Section 6, we extend our results by allowing the informed and uninformed segments to have heterogeneous service rewards and unit delay costs. This model is well suited to services with premium and regular customers when the category of a customer is predetermined by other exogenous factors. We find that system throughput and social welfare can still be unimodal in the fraction of informed customers.

2. Literature Review

The literature on the influence of real-time delay information on customer behavior dates back to Naor (1969). The author argues that in an observable service system, customer self-interested joining decisions, which ignore their negative externality on later arrivals, overload the system and result in a deviation from social optimality. Hassin and Haviv (2003) comprehensively summarize various extensions to Naor (1969). Hassin (1986) studies a revenue-maximizing server that has the option of completely suppressing the real-time information about the queue length. The author shows that when a revenue maximizer prefers to reveal the queue length, so does a social planner.

Customers may sometimes be unable to directly observe system states but have to rely on delay information offered by service providers. One example is delay announcements in call centers. Whitt (1999) argues that informing customers about anticipated delays can effectively reduce customer waiting times. Guo and Zipkin (2007) study the effects of providing information with different degrees of precision, i.e., no information, information on the queue length, and on the exact waiting time. They find that exact delay information may either improve or hurt social welfare because customers are not all equally patient. This finding is further strengthened by Guo and Zipkin (2009). These papers on delay announcements all implicitly assume that service providers offer truthful information. However, customers are often unable to verify the announced congestion information. Allon et al. (2011) model customers' strategic responses to the provider's unverifiable delay information and characterize equilibrium signaling languages that emerge between service providers and their customers. Allon and Bassamboo (2011) further reveal that delaying the announcements about waiting times can make the announced information more credible.

There is an emerging stream of literature on behavioral queues. Plambeck and Wang (2013) show how customers' lack of self-control and naiveté affect optimal pricing and scheduling in a service system. Huang et al. (2013) study canonical service models with boundedly rational customers. They find that for observable queues with endogenized pricing, bounded rationality results in a loss of revenue and welfare. Also under the notion of bounded rationality, Kremer and Debo (2016) predict and confirm with laboratory experiments that the high-quality firm's profit decreases in the fraction of informed customers when uninformed customers learn about the quality of a service from the queue length. Cui and Veeraraghavan (2016) study a queue that serves a pool of customers who may have arbitrarily misinformed, yet self-fulfilling, beliefs about the service rate. The authors show that revealing the service information to consumers can benefit revenues but may hurt individual or social welfare. Another stream of behavioral-queue research studies the herding effect. Veeraraghavan and Debo (2009, 2011) and Debo and Veeraraghavan (2014) study customer inferences about service quality through observation of the length of waiting lines, which may lead to herding in queues.

All the papers above assume that customer perceptions of delay information are homogeneous; i.e., either no one has access to the information or all receive the same types of information. However, as we argued before, it may be unrealistic to assume that all customers are aware of system congestion even though such information is available through many channels. By contrast to previous work, we consider customers' heterogeneous perceptions of delay information, i.e., only a fraction of customers can obtain the real-time queue length information. Our work focuses on the interaction between informed and uninformed customers and the resulting system performance.

Most relevant to our work is Hassin and Roet-Green (2017), which models rational customer decisions among three actions: join, balk or incur a *hassle cost* to inspect the queue length before making a join-or-balk decision. The authors prove the existence of a customer equilibrium strategy, which is effectively a randomization of the three possible actions. The authors show that the service provider can have a higher throughput if customers must incur a hassle cost to inspect the queue and therefore only a

fraction of customers are informed of the queue length. They also find that social welfare is maximized when the inspection is costless and thus all customers are informed.

By contrast, our base model can be adapted to account for the situation in which the service provider charges an *information fee* (considered a payment transfer between the provider and customers, unlike the hassle cost) and customers make rational decisions about whether to pay for being informed. Our results imply that charging an information fee to induce information heterogeneity can improve system throughput or social welfare. Our extension to rationalizing customers' uninformed behavior with an information fee may provide an alternative way of proving the equilibrium existence result in Hassin and Roet-Green (2017). In addition, we identify the equilibrium behavior of uninformed customers as the driving force of various comparative statics results and provide explanations.

Last, the information-ignorant behavior may also arise in other operations settings. Aflaki et al. (2015) find that a subset of forward-looking customers may intentionally choose to ignore the opportunity to search for more information. As a result, this rational ignorant behavior may increase the value of price commitment to the society as a whole.

3. Modeling Customer Join-or-Balk Decisions

A single-server facility expects a stream of customers who arrive one at a time according to a Poisson process with rate Λ. Customers are risk-neutral and are served on a first-come-first-served basis. The service time is an independent and identically distributed exponential random variable with mean $1/\mu$. We denote $\rho \equiv \Lambda/\mu$ as the offered load of the system. We assume that the admission fee to the facility is an irrelevant factor in a customer's join-or-balk decision and thus is assumed to be zero; as a result, the service provider's profit does not contribute to the social welfare. Such services may include border crossings, highways, and rides at Disney World.[2] Upon service completion, a customer receives a reward R. During the sojourn time in the system, a linear waiting cost with marginal rate c is incurred. All the parameters, Λ, μ, R, and c, are common knowledge. We further assume that the service reward is enough to offset the waiting cost when there is no line upon arrival, i.e., $R\mu \geq c$. If a customer chooses to balk, she receives zero utility. Moreover, we assume that customers do not renege.

There are two streams of customers. The *informed* stream checks the real-time information on the queue length, Q_t, and makes the join-or-balk decisions accordingly. The other *uninformed* stream ignores, or is unable to obtain, the real-time information. Thus, the uninformed customers have to rely on the expected queue length in deciding whether to join or balk. The fraction of informed customers in the whole population is denoted by an exogenous parameter $\gamma \in [0,1]$, which measures the prevalence of real-time information in the society. We also assume γ is common knowledge and will discuss this assumption in Subsection 3.2. We denote by λ_{I} and λ_{U} the arrival rates of informed and uninformed customer streams respectively. Then, we have

$$\lambda_{\mathrm{I}} \equiv \gamma\Lambda \quad \text{and} \quad \lambda_{\mathrm{U}} \equiv (1-\gamma)\Lambda.$$

We next discuss how the two customer streams make their joining decisions in equilibrium.

3.1. State-Dependent Decisions of Informed Customers

Informed customers observe the queue length in real time and make decisions that are contingent on the system state. After arriving and observing i customers in the system, including the one receiving service if any, an informed customer joins the queue if and only if the expected net value $R - c(i+1)/\mu \geq 0$, i.e., $i+1 \leq R\mu/c$. For simplicity of notation, $\lfloor x \rfloor$ denotes the largest integer that is less than or equal to x, and $\langle x \rangle \equiv x - \lfloor x \rfloor \in [0,1)$ denotes the *fractional part* of real number x. Therefore, there is a threshold

$$n \equiv \lfloor \nu \rfloor, \quad \text{where } \nu \equiv \frac{R\mu}{c} \geq 1,$$

such that an informed customer arriving at time t joins the queue if and only if she observes $Q_t < n$, and otherwise balks. In other words, $n-1$ is the maximum queue length beyond which joining the queue would lead to negative utility for an informed customer. Our model of informed customer behavior is in the same vein as those in observable queues (e.g., Naor 1969, Hassin 1986).

3.2. Mixed Strategy Equilibrium of Uninformed Customers

For simplicity, we follow the convention and restrict our attention to symmetric equilibrium strategies used by uninformed customers (see, e.g., Chapter 3 of Hassin and Haviv 2003). The uninformed customers do not know the real-time queue length. Their strategy can be described by a fraction $q \in [0,1]$, which represents the probability that each uninformed customer will join the queue, or equivalently, the proportion of all uninformed customers who will join the queue. Then, their expected net utility equals $R - cW(q)$, where $W(q)$ is the expected sojourn time for an uninformed customer under a joining probability q.

Note that the uninformed customers' knowledge of $W(q)$ requires their awareness of the fraction of informed customers, γ. This is a critical assumption

because it may sound odd that the uninformed customers are unable to access to the real-time queue length information, but they do know all system parameters, including the fraction of informed customers γ. On the one hand, this convenient but restrictive assumption allows us to focus on the impact of delay information heterogeneity on system performance.[3] On the other hand, it can be valid for some traditional service settings, such as the restaurant industry, where only drop-in customers can see the real-time queue length but other potential customers may not. It seems reasonable to assume that those customers who have to travel a long distance to see the queue can infer the fraction of drop-ins given the restaurant's characteristics, e.g., location and quality.[4]

We caution the assumption that the uninformed customers know the information level γ is a strong one. A less stylized model may incorporate the notion of bounded rationality and explicitly model the uninformed customers' learning of the expected delay as an outcome of their actions (i.e., the functional form of $W(q)$). For example, one may adopt the anecdotal reasoning framework in Huang and Chen (2015) to establish an asymptotic expected waiting time for the uninformed individuals. In addition, one may also use the idea in Cui and Veeraraghavan (2016) and assume that uninformed customers make their join-or-balk decisions under static but arbitrary beliefs about the information level γ. A potential problem of assuming static beliefs on γ is that uninformed customers' average experienced delays may differ from their "believed" expectations. One might resolve this issue by extending the idea to a dynamic setting where customers holding different beliefs may update theirs through repeated interactions with the system. These extensions require uninformed customers to perform elaborate trial-and-error computations and could also lead to information mis-calibration. Thus, these extensions that directly model the learning process of uninformed customers may or may not reach an equilibrium as what we characterize under the full information of γ. We leave these possible extensions to future research. In addition, we relax the common knowledge assumption on γ in Section 5 by allowing the service provider to control γ or having γ endogenized as the outcome of an information fee.

We next analytically characterize $W(q)$. Assume that the system is in the steady state, and let $p_i(q)$ denote the probability that there are i customers waiting in the queue when the joining probability of uninformed customers is q. The *collective* behavior of the informed and uninformed streams jointly determines the following balance equations of the process:

$$(\lambda_{\mathrm{I}} + q\lambda_{\mathrm{U}})p_i(q) = \mu \cdot p_{i+1}(q) \quad \text{if } 0 \le i < n, \tag{1}$$

$$q\lambda_{\mathrm{U}} \cdot p_i(q) = \mu \cdot p_{i+1}(q) \quad \text{if } i \ge n. \tag{2}$$

Thus, the probability mass function of the queue length can be written as:

$$p_i(q) = \begin{cases} (\rho_{\mathrm{C}}(q))^i p_0(q) & \text{if } 0 \le i < n, \\ (\rho_{\mathrm{C}}(q))^n (\rho_{\mathrm{U}}(q))^{i-n} p_0(q) & \text{if } i \ge n, \end{cases} \tag{3}$$

where, for notation convenience,

$$\rho_{\mathrm{C}}(q) \equiv \frac{\lambda_{\mathrm{I}} + q\lambda_{\mathrm{U}}}{\mu} \quad \text{and} \quad \rho_{\mathrm{U}}(q) \equiv \frac{q\lambda_{\mathrm{U}}}{\mu}. \tag{4}$$

(We may suppress the dependence of $\rho_{\mathrm{C}}(q)$ and $\rho_{\mathrm{U}}(q)$ on q to further simplify the notation.) Using $\sum_{i=0}^{\infty} p_i(q) = 1$, we can derive the idle probability

$$p_0(q) = \left(\sum_{i=0}^{n-1} (\rho_{\mathrm{C}})^i + \frac{(\rho_{\mathrm{C}})^n}{1-\rho_{\mathrm{U}}} \right)^{-1} = \left(\frac{1-(\rho_{\mathrm{C}})^n}{1-\rho_{\mathrm{C}}} + \frac{(\rho_{\mathrm{C}})^n}{1-\rho_{\mathrm{U}}} \right)^{-1}. \tag{5}$$

Therefore, the expected sojourn time W in the steady state is

$$W(q) = \sum_{i=0}^{\infty} \frac{i+1}{\mu} p_i(q) = \frac{p_0(q)}{\mu} \left[\frac{1-(\rho_{\mathrm{C}})^n}{1-\rho_{\mathrm{C}}} + \frac{\rho_{\mathrm{C}}}{(1-\rho_{\mathrm{C}})^2} + (\rho_{\mathrm{C}})^n \left(\frac{1-n}{1-\rho_{\mathrm{C}}} - \frac{1}{(1-\rho_{\mathrm{C}})^2} + \frac{1}{(1-\rho_{\mathrm{U}})^2} + \frac{n}{1-\rho_{\mathrm{U}}} \right) \right]. \tag{6}$$

Note that the system must be in the steady state when it reaches an equilibrium. If $0 < q^* < (=)1$, $cW(q^*) = (<)R$, and thus the expected queue length is finite. On the other hand, if $q^* = 0$, only informed customers will join the queue; as a result, the queue length never exceeds n.

If there were only uninformed customers, it is obvious that, ceteris paribus, the expected sojourn time $W(q)$ would increase as q becomes larger. However, we have two customer streams. As uninformed customers join the line more frequently, informed customers are less likely to join, which alleviates the congestion. The next lemma confirms that the former effect dominates the latter.

Lemma 1. *For any given $\gamma \in [0,1)$, the queue length $Q(q)$ in the steady state is stochastically increasing in q and thus the expected sojourn time $W(q)$ is strictly increasing in q.*

Following Hassin and Haviv (2003), we can determine the unique equilibrium joining probability of uninformed customers by the monotonicity of $W(q)$. The next proposition summarizes this.

Proposition 1. *Fix $\gamma \in [0,1)$. There exists a unique equilibrium joining strategy q^* for uninformed customers.*

(i) (Always Balk: No Participation) $q^* = 0$ *if and only if* $\mathbb{E}[Q(0)] \ge \nu$, *i.e.*, $cW(0) \ge R$.

(ii) (Always Join: Full Participation) $q^* = 1$ *if and only if* $\mathbb{E}[Q(1)] \le \nu$, *i.e.*, $cW(1) \le R$.

(iii) (Randomize between Balking and Joining: Partial Participation) $q^* \in (0,1)$ *must satisfy* $\mathbb{E}[Q(q^*)] = \nu$, *i.e.*, $cW(q^*) = R$, *if and only if* $\mathbb{E}[Q(0)] < \nu < \mathbb{E}[Q(1)]$, *i.e.*, $cW(0) < R < cW(1)$.

From the above proposition, we can further identify the system primitives, under which the uninformed customers always balk or join in equilibrium, by exploring the condition of $cW(0) \geq R$ or $cW(1) \leq R$, respectively. In the rest of the primitive space, uninformed customers randomize their decisions between joining and balking in equilibrium.

Corollary 1 (No Participation). *All uninformed customers always balk at the queue in equilibrium, i.e., $q^* = 0$, if and only if $1 \geq \gamma \geq \gamma_0^*(\rho, \nu) \equiv y^*(\nu)/\rho$, where $y^*(\nu) \geq 0$ is the unique solution to $n+1+1/(1-y)-(n+1)/(1-y^{n+1}) = \nu$.*[5]

Corollary 2 (Full Participation). *All uninformed customers always join the queue in equilibrium, i.e., $q^* = 1$, if and only if $1 \geq \gamma \geq \gamma_1^*(\rho, \nu) \equiv 1 - 1/\rho + (2/\rho)(\langle \nu \rangle + \sqrt{\langle \nu \rangle^2 + 4L(\rho, \nu)})^{-1}$, where $L(\rho, \nu) \equiv (\langle \nu \rangle(\rho - 1)\rho^n + \nu - \nu\rho + \rho^n - 1)/((1-\rho)^2\rho^n) = (\nu - \langle \nu \rangle \rho^n - \sum_{i=0}^{n-1} \rho^i)/((1-\rho)\rho^n) \geq 0$ and $\langle \nu \rangle \equiv \nu - n$.*[6]

Corollaries 1 and 2 show that uninformed customers will always join or balk when the information level is above a certain threshold. Nevertheless, depending on the system offered load, only one threshold, $\gamma_0^*(\rho, \nu)$ or $\gamma_1^*(\rho, \nu)$, can exist in the range of $[0, 1]$ for a given set of system primitives. The next proposition describes how the offered load ρ and information level γ jointly determine the equilibrium joining behavior of uninformed customers.

Theorem 1 (Equilibrium Strategy of Uninformed Customers). *For given ρ and ν, define $\underline{\rho} \equiv 1 - 1/\nu$ and $\bar{\rho} \equiv y^*(\nu)$, respectively. The equilibrium joining probability q^* of uninformed customers depends on ρ and γ in the following way*:

(i) (Always Full Participation) *If $0 \leq \rho < \underline{\rho}$, $q^* = 1$ for all $0 \leq \gamma \leq 1$.*

(ii) (Partial to Full Participation) *If $\underline{\rho} \leq \rho \leq \bar{\rho}$, $q^* \neq 0$ for all $0 \leq \gamma \leq 1$. In particular, $0 < q^* < 1$ for $0 \leq \gamma < \gamma_1^*(\rho, \nu)$ and $q^* = 1$ for $\gamma_1^*(\rho, \nu) \leq \gamma \leq 1$, where $\gamma_1^*(\rho, \nu) \in [0, 1]$.*

(iii) (Partial to No Participation) *If $\rho > \bar{\rho}$, $q^* \neq 1$ for all $0 \leq \gamma \leq 1$. In particular, $0 < q^* < 1$ for $0 \leq \gamma < \gamma_0^*(\rho, \nu)$ and $q^* = 0$ for $\gamma_0^*(\rho, \nu) \leq \gamma \leq 1$, where $\gamma_0^*(\rho, \nu) \in [0, 1]$.*

We next use Figure 1 to illustrate the results in Theorem 1. The vertical axis represents the system offered load $\rho = \Lambda/\mu$, and the horizontal axis is $\nu = R\mu/c$, which ranges from n inclusive to $n+1$ exclusive, for some $n \geq 1$. The lower dashed and upper solid curves correspond to the two thresholds $\underline{\rho}$ and $\bar{\rho}$ respectively. Moreover, it can be shown that the solid curve always stays above the dashed one and $\bar{\rho}$ approaches infinity when ν tends to $n+1$. The two offered load thresholds $\underline{\rho}(\nu)$ and $\bar{\rho}(\nu)$ divide uninformed customer equilibrium strategies into three types in the primitive space of (ρ, ν). We discuss each below.

In the area below $\underline{\rho}$, the offered load is very low and the expected sojourn time for uninformed customers is very short. Even if no one observes the queue length, i.e., $\gamma = 0$, all uninformed customers choose to join the queue, i.e., $q^* = 1$. As the information level γ increases by an infinitesimal amount, a small proportion of customers change from uninformed to informed. These converted customers now join the queue only if they observe that $Q_t < n$, rather than join definitely as before. System congestion is hence slightly alleviated. The reduced delay, on the one hand, reinforces the remaining uninformed customers' incentives to join, with the result that they all still join the queue as the information level γ grows. This also intuitively justifies the validity of Corollary 2. On the other hand, the slightly reduced congestion also increases the probability that informed customers will see a short queue. Therefore, the chance that an informed customer joins the line slowly increases in γ when $q^* = 1$. Figure 2(a) depicts q^* as a function of γ in such a scenario when $\rho = \Lambda/\mu = 0.5$, the service reward $R = 4$, and marginal waiting cost $c = 1$. In this case, uninformed customers always join the queue in equilibrium regardless of the information level as shown in Theorem 1(i), and informed customers' probability of joining rises slightly as γ increases.

As the offered load increases to intermediate levels between the dashed and solid lines in Figure 1, a large fraction or all of the uninformed customers join the queue in equilibrium depending on the information

Figure 1. (Color online) Illustration for Equilibrium Behavior of Uninformed Customers

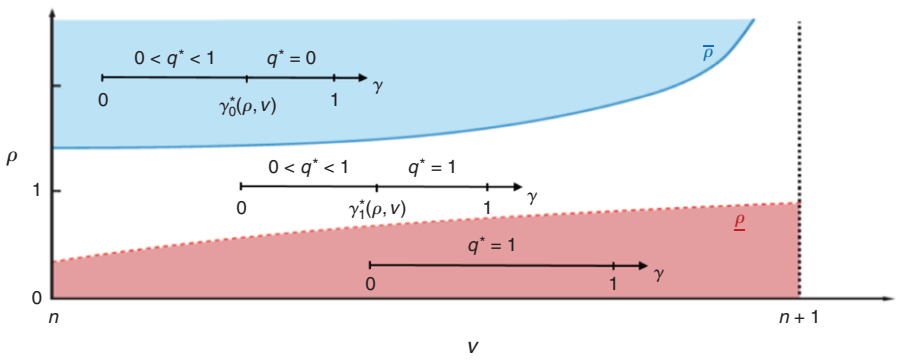
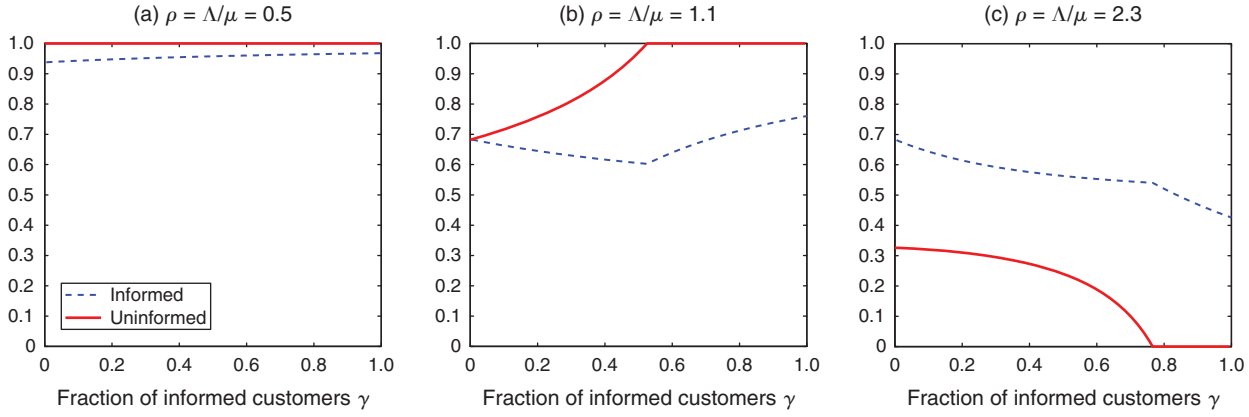


Figure 2. (Color online) Joining Probabilities of Informed and Uninformed Customers ($\mu = 1$, $R = 4$, and $c = 1$)



level γ (see Theorem 1(ii)). We first consider the case in which almost all customers are informed, e.g., when $\gamma \approx 1$. Because of the relatively modest offered load, uninformed individuals still have a strong incentive to join. In fact, they all join, i.e., $q^* = 1$. Moreover, the likelihood that an informed customer will join the line would decline slightly if a small number of informed customers became uninformed, because the converted customers who used to join only when the queue length was less than n will now undoubtedly join. This phenomenon is depicted in Figure 2(b) for $\gamma \geq 0.51$, when the offered load $\rho = 1.1$. As the number of uninformed individuals rises further, i.e., as $\gamma \to 0^+$, the majority of the customers who are lining up are uninformed and must seek service less often to cope with the increasing negative externalities from their uninformed peers. Thus, q^* decreases as $\gamma \to 0^+$, which tends to increase the likelihood that the queue will be shorter than n because of the modest offered load and hence the probability that an informed customer will join. Figure 2(b) also displays the dynamics when $0 \leq \gamma < 0.51$.

As the offered load rises to high levels, i.e., above the solid curve in Figure 1, only a fraction or none of the uninformed customers join the queue in equilibrium, depending on the information level γ (see Theorem 1(iii)). In this case, the system expects enormous customer volume. With no knowledge of the real-time queue length, uninformed customers expect a long line and have very little incentive to join. With such a large customer volume, the advantage of real-time congestion is clear: Any available spot with fewer than n people ahead will be quickly taken. As the fraction of informed customers increases, this effect becomes more salient and the queue becomes even longer. Informed customers observe a short queue less often and have to balk more often as γ increases to 1. For uninformed customers, the incentive to join the line diminishes in γ until it vanishes. For the same reason, a further increment in the fraction of informed customers beyond the vanishing point where q^* hits 0 only causes those appealing spots to be occupied even faster and the queue to be even longer. Hence, q^* remains at 0 after the information level exceeds $\gamma_0^*(\rho, \nu)$, as shown in Corollary 1. Figure 2(c) displays such dynamics when the offered load $\rho = 2.3$.

4. Effects of Heterogeneous Information

To answer the questions raised in the Introduction, we investigate how a marginal increase in the information level would affect system efficiency. The first central step is to define the notion of efficiency. To an engineer, throughput, which implies the use of the server, might be the relevant efficiency measure. To an economist, social welfare, which includes the overall economic benefit, may be a more suitable measure. Therefore, we consider the system throughput and social welfare as our performance measures in this paper.

The system throughput, denoted by λ, consists of the effective arrival rates of informed and uninformed streams. That is,

$$\lambda(q) = \left(\sum_{i=0}^{n-1} p_i(q)\right)\lambda_{\mathrm{I}} + q\lambda_{\mathrm{U}}. \tag{7}$$

Likewise, social welfare is composed of contributions from both segments. Let $S_{\mathrm{I}}(q)$ and $S_{\mathrm{U}}(q)$ be the net utilities accruing to informed and uninformed customers in unit time, respectively. We have

$$S_{\mathrm{I}}(q) = \left[\sum_{i=0}^{n-1} p_i(q)\left(R - c\frac{i+1}{\mu}\right)\right]\gamma\Lambda \quad \text{and}$$

$$S_{\mathrm{U}}(q) = \left[q\sum_{i=0}^{\infty} p_i(q)\left(R - c\frac{i+1}{\mu}\right)\right](1-\gamma)\Lambda$$

$$= q(R - cW(q))(1-\gamma)\Lambda.$$

Social welfare is thus the sum of all customer net utilities, i.e.,

$$S(q) = S_{\mathrm{I}}(q) + S_{\mathrm{U}}(q). \tag{8}$$

Although informed customers use the same state-dependent threshold strategy as they would in an

observable queue (see Naor 1969), their contributions to the effective arrival rate and social welfare also hinge on the uninformed customers' strategy q through $p_i(q)$, $i = 0, \ldots, n-1$. On the one hand, this interaction, which is not captured by the observable model or the unobservable one, implies the decisive role of the uninformed customers' equilibrium strategy q^* in a general circumstance. On the other hand, it also increases the technical difficulty of the analysis. One may be tempted to seek a general closed-form expression for the equilibrium joining probability q^* of uninformed customers as a function of γ and conduct comparative statics of $\lambda(q^*(\gamma))$ and $S(q^*(\gamma))$ on the information level γ. However, it is a daunting, if not impossible, task. That is because the equilibrium joining probability $q^* \in (0,1)$ is characterized by a high-order polynomial equation: $cW(q) = R$, where $W(q)$ is specified in Equation (6). According to the Abel-Ruffini Theorem, this type of equation in general has no algebraic solution in radicals. Although the insolvability of q^* hinders a direct approach to analyzing the system, we will show monotonicity properties of equilibrium throughput and social welfare with respect to the information level γ by taking an indirect approach. Somewhat surprisingly, their monotonicity properties are uniquely determined by the *type* of equilibrium joining strategy that uninformed customers adopt, i.e., always balk (i.e., $q^* = 0$), always join (i.e., $q^* = 1$) or randomize between balking and joining (i.e., $0 < q^* < 1$), which is an outcome of interactions between informed and uninformed customers.

4.1. Throughput

Chen and Frank (2004) compare the throughput in the observable (i.e., $\gamma = 1$) and unobservable (i.e., $\gamma = 0$) queues. They demonstrate that there is a unique critical level ρ^* such that if $\rho > \rho^*$, the throughput of the observable queue is more than that of the unobservable queue; the converse holds if $\rho < \rho^*$. Although their result provides an answer to the comparison of $\lambda(q^*(\gamma = 0))$ and $\lambda(q^*(\gamma = 1))$, it does not reveal the marginal effect of information on the equilibrium throughput $\lambda(q^*(\gamma))$ in general for $\gamma \in [0,1]$. In this subsection, we characterize the monotonicity of the system equilibrium throughput λ in the information level γ.

Note that it is difficult to directly analyze the throughput formula $\lambda(q)$ in (7). However, if we take a summation of all system states in (1) and (2), we observe that

$$\sum_{i=0}^{n-1}(\lambda_{\mathrm{I}} + q\lambda_{\mathrm{U}})p_i(q) + \sum_{i=n}^{\infty} q\lambda_{\mathrm{U}}p_i(q) = \mu\sum_{i=0}^{\infty} p_{i+1}(q)$$

$$\iff \lambda(q) = \left(\sum_{i=0}^{n-1} p_i(q)\right)\lambda_{\mathrm{I}} + q\lambda_{\mathrm{U}} = \mu(1 - p_0(q)).$$

In other words, the system throughput equals the service rate minus the vacant capacity due to idleness. Therefore, we can explore the monotonicity of the equilibrium throughput via the probability of idleness in equilibrium.

Lemma 2. *If $q^* \in [0,1)$, the probability $p_0(q^*(\gamma))$ that the server is idle strictly decreases in γ.*

Lemma 2 indicates that as long as uninformed customers do not take the full-participation strategy in equilibrium, the server is less likely to be idle as the real-time congestion information becomes more prevalent. This is in sharp contrast to the case of $q^* = 1$, where $p_0(q^*(\gamma) = 1)$ increases in γ. Recall from Corollary 2 that q^* is at 1 when γ hits $\gamma_1^*(\rho, \nu)$. A further increase in γ turns a fraction of uninformed customers who used to join definitely into informed customers who line up only if the queue is shorter than n. As a result, the number of customers who actually join decreases as γ increases and the server is more likely to be idle. The next theorem summarizes the effects of changing server idleness on system throughput due to growing information prevalence.

Theorem 2 (Comparative Statics of Throughput). (i) *If $0 \le q^* < 1$, the throughput $\lambda(q^*)$ is strictly increasing in γ.*

(ii) *If $q^* = 1$, the throughput $\lambda(q^*)$ is strictly decreasing in γ.*

Theorem 2 says that system throughput benefits from growing information prevalence unless all uninformed customers choose to join the queue in equilibrium. However, the negative effect of information on throughput can occur over a large range. We illustrate that with an example.

Example 1. Consider a system with the offered load ρ that is close to 1. From Corollary 2, $\lim_{\rho\to 1} L(\rho, \nu) = \langle\nu\rangle n + n(n-1)/2$ by L'Hôpital's rule and hence, $\lim_{\rho\to 1} \gamma_1^*(\rho, \nu) \le 2/\sqrt{2n(n-1)}$, which can be further bounded by $1/\sqrt{3} \approx 57.7\%$ when $n - 1 \ge 2$.

Given that uninformed customers always choose to join the queue when $\gamma \ge \gamma_1^*(\rho, \nu)$ (see Corollary 2), Theorem 2 implies that if informed customers have a joining threshold no less than 2, throughput will always suffer if there are more than 58% of informed customers in the population. This simple example clearly shows that system throughput can easily suffer from real-time congestion information if too many customers are informed. This phenomenon is a result of equilibrium responses by uninformed customers to the heterogeneous availability of information. Hence, it is not covered by the conventional observable and unobservable frameworks. □

We next explain the intuition of Theorem 2 through its connection to Lemma 2. As we have established, maximizing the throughput is equivalent to minimizing the server's probability of being idle, $p_0(q^*)$. In principle, there are two reasons for idleness, i.e., inadequate customer arrivals (i.e., low arrival rate relative

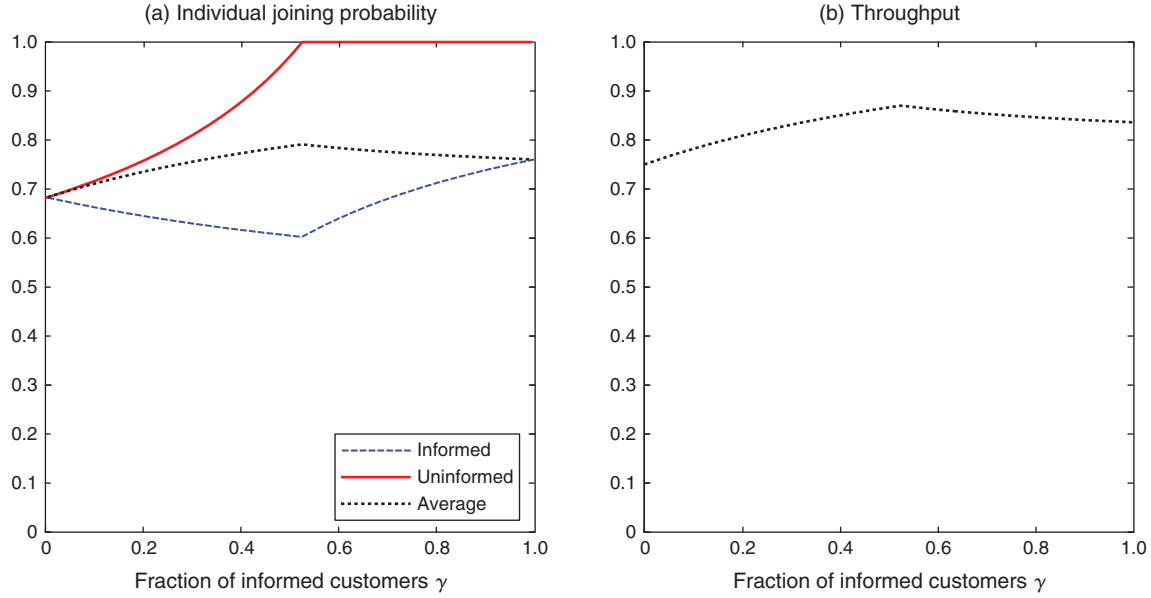
to the service speed) and an intertemporal mismatch between capacity and demand, i.e., the mean and the variability effects. To illustrate, consider an extreme case in which the system has deterministic service and interarrival times. For such a system, if the offered load is small, say strictly less than one, the server inevitably becomes idle upon completion of a job and has to wait a while for the next customer. Obviously, this type of idleness is caused by the low customer arrival rate relative to the service time. Thus, we refer to it as the mean effect. For the same deterministic system, if the offered load is large, say bigger than or equal to one, the server is fully used and the throughput equals the service rate. However, any variability in service or interarrival times can occasionally cause the server to be idle, referred to as the variability effect. The above discussion also suggests that the dominant effect in causing idleness depends on the offered load. We elaborate below.

When the offered load is relatively low, the customer arrival rate is low, and thus the server is very likely to complete all tasks before another customer arrives. Hence, having too few arrivals is the first-order effect that gives rise to idleness. To improve throughput, the provider has to increase the average probability that an individual customer will join the line. This rationale can also be seen mathematically. Rearranging (7), we write the throughput as

$$\lambda(q^*(\gamma)) = \left(\sum_{i=0}^{n-1} p_i(q^*(\gamma))\gamma + q^*(\gamma)\cdot(1-\gamma)\right)\Lambda,$$

where the term in parentheses represents the average joining probability of the entire population. We next examine the role of information in motivating customers to join the queue. Note that when the offered load is low, the queue is expected to be short regardless of the information level γ. Although a slight change in γ does affect the probability that the queue will be shorter than n and thus the chance that an informed individual will join the queue, such an effect is very marginal because of the low offered load. Therefore, an informed customer's likelihood of joining is not very sensitive to information level changes, upward or downward. By contrast, uninformed customers are more sensitive to the change in information prevalence. As more uninformed individuals become informed, the number of uninformed customers clearly decreases. Moreover, those who used to be uninformed but become informed also join less often because they now anticipate positive, instead of zero, net utility to justify their participation. These two effects together give the remaining uninformed customers a stronger incentive to join. Consequently, as information becomes more prevalent, uninformed customers are much more motivated to join the queue. Note that their enthusiasm reduces the chance that the informed customers will join. Still, as we argued before, this effect is marginal because of the low offered load. Combining both segments, we can see that the average customer joining probability rises as a function of the information level γ, with the uninformed segment being more strongly motivated and the informed segment being slightly discouraged. The dotted curve in Figure 3(a) shows the weighted average joining probability of the entire customer pool. As we see, it increases to $\gamma = 0.51$, at which point uninformed customers start to join definitely, i.e., $q^* = 1$. Additional information

Figure 3. (Color online) Example: $\mu = 1$, $R = 4$, $c = 1$, and $\Lambda = 1.1$



prevalence is unable to further stimulate uninformed customers; hence, it loses its beneficial effect in motivating them. As more uninformed customers become informed after q^* hits 1, they join less than when they were uninformed. Consequently, the average joining probability decreases, as shown by the dotted curve in Figure 3(a) beyond $\gamma = 0.51$. Figure 3(b) displays the throughput as a function of γ. The pattern in Figure 3(b) exactly matches that of the average joining probability in Figure 3(a).

When the offered load is high enough, there are ample customer arrivals relative to the service speed. An intertemporal mismatch between capacity and arrivals due to system variability becomes the primary reason for server idleness. In this situation, increasing the number of informed customers has two effects. On the one hand, as more customers become informed, the desirable positions with fewer than n persons waiting ahead are taken more instantaneously. On the other hand, with more informed customers, fewer customers make uninformed decisions and abandon a short queue. Both effects reduce the likelihood of server idleness, thus improving the throughput.

In summary, increasing the availability of information improves system throughput unless all the uninformed customers join the queue. However, the reason for the phenomenon varies for different offered loads. Under a low offered load, growing information prevalence improves the average joining probability of the entire customer pool and also the throughput. However, the throughput declines if uninformed customers cannot be further motivated when all of them have already chosen to join. By contrast, if the offered load is high, growing information prevalence helps minimize the risk of idleness by more effectively matching capacity with demand.

Our reasoning can easily explain the findings of Chen and Frank (2004). When $\rho < \rho^*$, the offered load is low. The service provider wants to increase the probability that each customer will join the queue. In that case, an unobservable queue is preferable since customers will only join under positive utility if they are informed, but will tolerate zero utility if they are uninformed. When $\rho > \rho^*$, the offered load is high. Minimizing mismatch due to uncertainty through information disclosure is an effective tactic. Thus, an observable queue is preferable.

Recall that the equilibrium strategy of uninformed customers depends on the offered load according to Theorem 1. Thus, the following result, as a direct corollary of Theorems 1 and 2, specifies the effect of growing information prevalence on the throughput in the primitive space (ρ, ν).

Corollary 3. *For given ρ and ν, define $\underline{\rho} \equiv 1 - 1/\nu$ and $\bar{\rho} \equiv y^*(\nu)$, respectively. Then,*

(i) *if $0 < \rho < \underline{\rho}$ the throughput $\lambda(q^*)$ is strictly decreasing in γ;*

(ii) *if $\underline{\rho} \le \rho \le \bar{\rho}$, the throughput $\lambda(q^*)$ is strictly increasing in $\gamma \in [0, \gamma_1^*(\rho, \nu))$ and is strictly decreasing in $\gamma \in [\gamma_1^*(\rho, \nu), 1]$;*

(iii) *if $\rho > \bar{\rho}$, the throughput $\lambda(q^*)$ is strictly increasing in γ.*

Corollary 3 reveals that heterogeneity of information about the real-time queue length can effectively improve the system throughput, except for the case in which the offered load is sufficiently low or high. We refer to Figure 1 to elaborate. Case (i), where information always hurts the throughput, corresponds to the area below the dashed line. Case (iii), where information always benefits the throughput, corresponds to the area above the solid line. The more intriguing case (ii), where the equilibrium throughput $\lambda(q^*)$ is unimodal in γ, corresponds to the intermediate area between the two lines. In this area, there is always an intermediate information level $\gamma_1^*(\rho, \nu)$ that maximizes system throughput. Therefore, the system throughput with heterogeneous congestion information outperforms its counterparts where all customers are equally informed or uninformed.

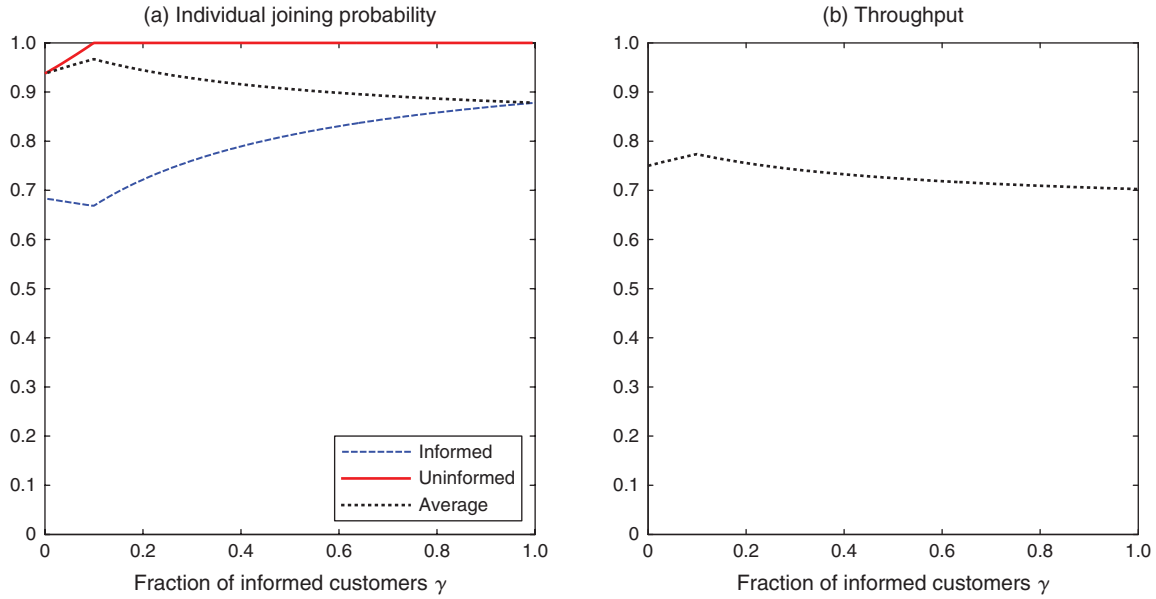
The comparison of unobservable and observable queues in Chen and Frank (2004) is a special case of Corollary 3 for comparing $\gamma = 0$ and $\gamma = 1$. Our result shows that the unobservable and observable queues are preferred in cases (i) and (iii), respectively. Therefore, Corollary 3 implies that the critical level ρ^* in Chen and Frank (2004) must lie between $\underline{\rho}$ and $\bar{\rho}$. Moreover, as illustrations for case (ii), Figure 3 shows that the observable queue is preferable to the unobservable counterpart when $\rho = 1.1$, whereas Figure 4 shows that the opposite is true when $\rho = 0.8$. Hence, the critical level ρ^* in Chen and Frank (2004) is between 0.8 and 1.1 for the given set of parameters.

4.2. Social Welfare

In this subsection, we focus on social welfare. The existing literature argues that real-time congestion information is efficient in improving social welfare because it helps customers make efficient decisions: They do not join a long queue and do not balk at a short one. Thus, it is believed that delay information should be disclosed for the benefit of society. Thanks to advances in information technology, it is easier than ever to obtain all kinds of congestion information for public facilities, e.g., border services and highways. Is it really true that all customers being informed always maximizes social welfare? We answer that question in this subsection. We first discuss the influence of the increasing availability of information on individual net utility and then consider social welfare as a whole.

For ease of exposition, we denote the individual net utility of informed and uninformed customers by $\bar{S}_I(q)$

Figure 4. (Color online) Example: $\mu = 1$, $R = 4$, $c = 1$, and $\Lambda = 0.8$



and $\bar{S}_{\mathrm{U}}(q)$, respectively. Specifically,

$$\bar{S}_{\mathrm{I}}(q) = \sum_{i=0}^{n-1} p_i(q)\left(R - c\frac{i+1}{\mu}\right) \quad \text{and}$$

$$\bar{S}_{\mathrm{U}}(q) = q\sum_{i=0}^{\infty} p_i(q)\left(R - c\frac{i+1}{\mu}\right) = q(R - cW(q)).$$

By Proposition 1, an individual uninformed customer receives a nonzero utility only if $q^* = 1$ in equilibrium. An informed customer earns not only a nonnegative utility but also a higher utility than an uninformed customer at any information level. It turns out that the monotonicities of $\bar{S}_{\mathrm{I}}$ and $\bar{S}_{\mathrm{U}}$ in the information level are also uniquely determined by the *type* of uninformed customers' equilibrium actions.

Theorem 3 (Comparative Statics of Individual Welfare). (i) *$\bar{S}_{\mathrm{I}}(q^*)$ is strictly decreasing in γ if $0 \le q^* < 1$ and is strictly increasing in γ if $q^* = 1$.*

(ii) *$\bar{S}_{\mathrm{U}}(q^*) = 0$ if $0 \le q^* < 1$ and $\bar{S}_{\mathrm{U}}(q^*)$ is strictly increasing in γ if $q^* = 1$.*

The results in the above theorem can be considered as implications of Theorem 2. When $0 \le q^* < 1$, the system throughput increases in the information level γ by Theorem 2. Hence, the system is expected to be more congested, a situation that erodes the net utility of each informed individual. Yet when $q^* = 1$, as γ increases, the throughput decreases and system congestion is relieved. Hence, informed and uninformed individuals obtain more net utility.

We next consider the total consumer net utility, i.e., social welfare. By (8), we have

$$S(q) = S_{\mathrm{I}}(q) + S_{\mathrm{U}}(q) = \bar{S}_{\mathrm{I}}(q)\cdot\gamma\Lambda + \bar{S}_{\mathrm{U}}(q)\cdot(1-\gamma)\Lambda.$$

In the case of $q^* \in [0,1)$, since $S_{\mathrm{U}}(q^*) = 0$, social welfare is simply the total utility of informed customers. Although the individual net utility of informed customers decreases in γ for $q^* \in [0,1)$, the number of informed customers increases. The next result shows that the availability of information improves social welfare unless no uninformed customers are interested in the service, i.e., $q^* = 0$.

Theorem 4 (Comparative Statics of Social Welfare). (i) *If $q^* = 0$, the social welfare $S_{\mathrm{I}}(q^*) + S_{\mathrm{U}}(q^*)$ is strictly increasing in γ for $1 \le \nu < 2$ and is strictly decreasing in γ for $\nu \ge 2$.*

(ii) *If $0 < q^* \le 1$, the social welfare $S_{\mathrm{I}}(q^*) + S_{\mathrm{U}}(q^*)$ is strictly increasing in γ.*

Theorem 4 first confirms the conventional wisdom that, in general, real-time congestion information can efficiently match the potential available capacity of a system with customer demand. That is, with congestion information, customers can join the service when they observe a short queue, which signals forthcoming availability of the server. We may therefore expect that real-time information always improves social welfare. However, a potentially negative effect due to the (negative) queueing externality may also arise from the disclosure of congestion information, as Theorem 3(i) reveals. That is, growing information prevalence may also result in constantly declining individual utility, specifically when the system faces a high offered load. As the fraction of informed customers increases, a greater number of informed customers are competing with one another for the desirable positions with fewer than n customers ahead of them. Because of the high offered load and the efficiency of information in matching waiting slots with demand, such positions are quickly taken when they are available. Therefore,

as γ increases, informed customers are expected to see a longer queue upon arrival and to be more likely to balk, both of which reduce their individual net utility.

The diminishing individual net utility of informed customers would not cause a loss of social welfare as long as uninformed customers were still interested in the service, i.e., when $q^* > 0$. That is because, when the offered load and the information level are high, uninformed customers, given the disadvantage of being unable to make informed decisions contingently, compromise by joining the queue infrequently since they expect a long line. The low incentive for them to join the line helps reduce the congestion. When an even larger portion of the high-volume customers become informed, uninformed customers have an even lower incentive to join the line. This declining interest in the service frees up the tight capacity for informed customers, who earn higher utility than uninformed customers. More important, the reduced engagement of uninformed customers alleviates intensified competition among informed customers for the appealing positions and thus prevents informed customers' joining probability and individual net utility from quickly declining. Nonetheless, after all uninformed customers lose interest in the service and choose to balk, i.e., when $q^* = 0$, growing information prevalence only increases the fraction of informed customers, whose individual net utility therefore decreases faster when $q^* = 0$, than when $q^* > 0$, without the compromise made by uninformed customers. As a result, if no uninformed customers consider joining, social welfare starts to decline as real-time delay information is more prevalent.

Figure 5 illustrates the patterns with $\rho = 2.3$. For informed customers, their joining probability and individual net utility decline with the information level γ. However, when $q^* > 0$, both metrics noticeably decrease more slowly than when $q^* = 0$. As we have discussed, this difference can be explained by uninformed customers' disincentives to join. This declining incentive to join, together with the declining number of the uninformed customers, helps the system use its tight capacity to serve more informed customers who can yield higher welfare. These effects no longer exist when the uninformed customers completely refrain from joining.

Combining Theorems 1 and 4, we can identify the monotonicity properties of the social welfare under different offered loads.

Corollary 4. *For $1 \le \nu < 2$, the social welfare is strictly increasing in γ. For $\nu \ge 2$,*

(i) *if $0 \le \rho \le \bar{\rho} \equiv y^*(\nu)$, the equilibrium social welfare is strictly increasing in γ;*

(ii) *if $\rho > \bar{\rho}$, the equilibrium social welfare is strictly increasing in $\gamma \in [0, \gamma_0^*(\rho, \nu))$ and is decreasing in $\gamma \in [\gamma_0^*(\rho, \nu), 1]$.*

Theorem 4 and Corollary 4 have two implications. First, full transparency in congestion information may not achieve the highest efficiency. Specifically, a highly-loaded system yields the most social welfare if some customers are uninformed. Second, even if information is not at the optimal level $\gamma_0^*(\rho, \nu)$, the social welfare under heterogeneous information structures can still outperform a completely observable system for a large range of information levels. For instance, social welfare for any $\gamma \in (0.38, 1)$ is higher than when $\gamma = 1$ in the example shown in Figure 5.

As an exception, if customers have a very low service reward R such that the informed customers join the queue only when it is empty, in which case $1 \le \nu < 2$, information always helps, even when $q^* = 0$. From the perspective of a server, the line temporarily holds waiting customers and supplies the server with customers upon completion of a job. If $q^* = 0$, even uninformed customers do not join a queue. As a result, no one intends to queue at all. After completing a service job, the server must remain idle and wait to resume creating surplus until the next customer arrives. Therefore, the system has to rely on information ubiquity

Figure 5. (Color online) Example: $\mu = 1$, $R = 4$, $c = 1$, and $\Lambda = 2.3$

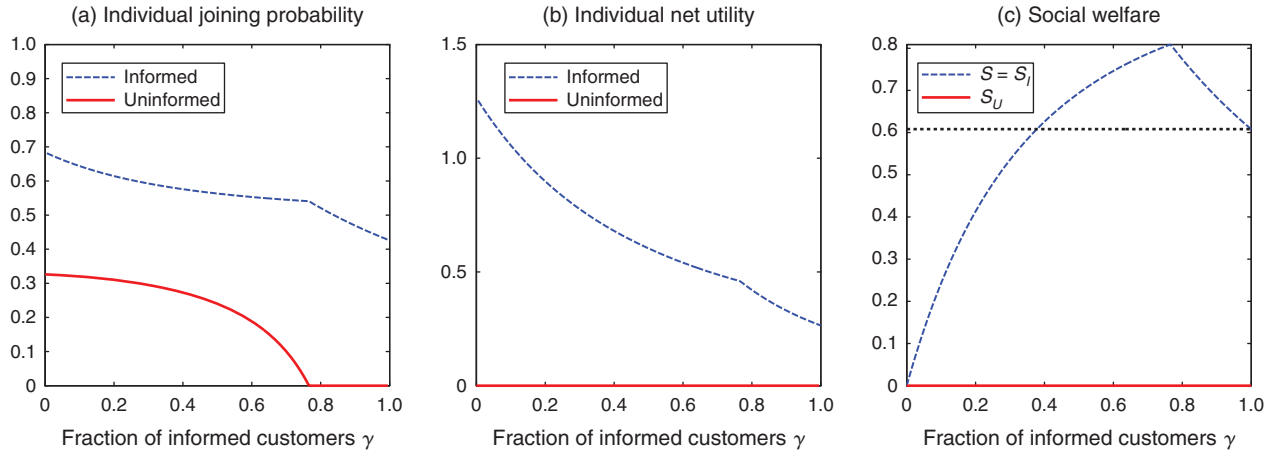


Table 1. Comparative Statics on γ

	$q^* = 0$	$0 < q^* < 1$	$q^* = 1$
Throughput	↑	↑	↓
Individual welfare of informed customers	↓	↓	↑
Individual welfare of uninformed customers	0	0	↑
Social welfare	↓[a]	↑	↑

[a]↑ for $\nu \in [1, 2)$.

to secure a customer as soon as its capacity is available. Behavior that is so different from the general case, when informed customers are only willing to be in the system alone, is also observed in Hassin (1986).

4.3. Summary

Table 1 summarizes the effects of growing information prevalence on various performance measures when γ is exogenous. The effects depend on the type of equilibrium joining strategy of uninformed customers, specified by their equilibrium joining probability q^*. One can easily visualize the effects of information prevalence and the optimal information levels in the primitive space (ρ, ν), illustrated in Figure 1, by combining Theorem 1 and Table 1.

If throughput is the focal performance measure, the service provider should reveal the queue length and encourage information dissemination when the offered load is high enough (the top area in Figure 1) and conceal it when the offered load is low enough (the bottom area in Figure 1). If the offered load is in an intermediate range (the middle area in Figure 1), it is optimal to have a segment of uninformed customers or reveal the real-time information only to a fraction of customers. However, if social welfare is the focal performance measure, the service provider should reveal the queue length information and encourage its dissemination when the offered load is relatively small (the areas below the solid line in Figure 1). In other situations (the area above the solid line, i.e., the top area, in Figure 1), it is optimal to have a segment of uninformed customers or, equivalently, to hide the real-time congestion information from certain customers. In the next section, we further discuss control of information heterogeneity.

5. Achieving the Optimal Information Level

Our base model gives an optimistic view of the effects of information heterogeneity. The fact that some customers do not have real-time delay information indeed helps system throughput when the offered load is relatively low and improves the social welfare when the offered load is relatively high. However, our analysis so far assumes that the information level is exogenous. In this section, we discuss two strategies that allow the service provider to achieve optimal outcomes. The first strategy is to randomize or control the granularity of delay information (real-time or expected) offered to customers if the service provider is in charge of the disclosure of such information, such as announcements of delays at call centers or emergency rooms. The second strategy results in the optimal fraction of informed customers by charging a fee for obtaining real-time congestion information.

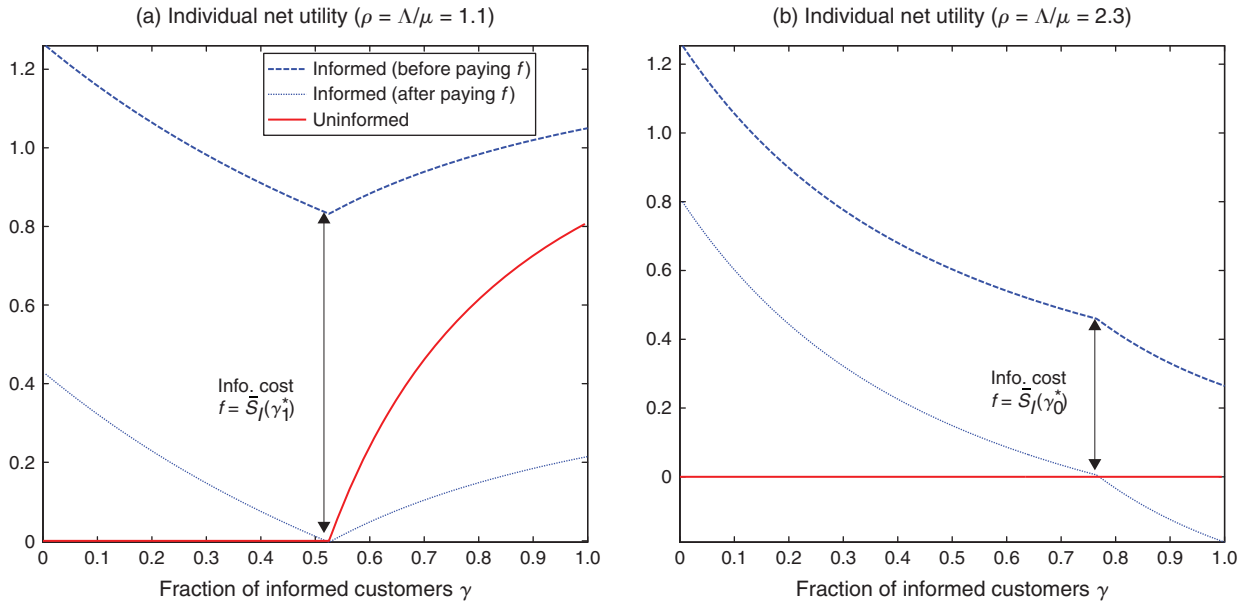
5.1. Disclosing Heterogeneous Delay Information to Customers

Given the benefit of information heterogeneity, service providers may consider achieving higher performance by deliberately creating such heterogeneity. For example, call centers can determine the desired number of informed and uninformed customers according to system attributes, e.g., total arrival and service rates. While offering real-time delay information to some customers, they may also estimate and announce only the expected delay to others. Service providers may also randomize the information between real-time and expected delay for all visitors. Specifically, a service provider can simply implement the optimal solution by revealing the real-time information with probability γ^* and the expected delay[7] with probability $1 - \gamma^*$, where γ^* is the optimal fraction of informed customers. This tactic allows the provider to influence customers' joining behavior through congestion information of varying accuracy. In addition, it does not require the uninformed stream to adjust their joining behavior through trial and error to reach the equilibrium, which may be unrealistic in certain circumstances, e.g., emergency room visits.

5.2. Endogenizing Information Levels Through Information Fee

We now discuss the use of a fee to regulate self-interested customer decisions. Note that information ignorance seems *irrational* if access to the information is free. Uninformed customers who do not obtain real-time information always earn less utility than informed customers. Therefore, if congestion information is free and convenient, a rational, uninformed customer has every incentive to learn how long the queue is. Consequently, all customers would choose to be informed in equilibrium under self-interested rational choices. Next we discuss how the service provider can achieve the optimal information level by charging an information fee when customers are fully rational. For this purpose, we temporarily assume in this subsection that all customers exhibit self-interested, utility-maximization behavior and that they know the system parameters Λ, μ, R, and c.

Figure 6. (Color online) Illustration of Information Fee ($\mu = 1$, $R = 4$, and $c = 1$).



5.2.1. Inducing the Optimal Throughput. The optimal information level can be easily induced when the offered load is very high or very low. Recall that if $\rho > \bar{\rho}$, the system throughput is maximized when all customers are informed, i.e., $\gamma = 1$. Therefore, with congestion information revealed by the provider, the self-interested choice by all customers to be informed will lead to an equilibrium that is sustained as the system optimum. If $\rho < \underline{\rho}$, no one being informed leads to the maximum throughput. Thus, the provider can simply conceal real-time congestion information such that no customers can be informed.

In the case where the offered load is in an intermediate range, i.e., when $\underline{\rho} \le \rho \le \bar{\rho}$, the system throughput is maximized at $\gamma = \gamma_1^*$ according to Corollary 3. The optimal information level is not achievable through decentralized actions by customers under queue-length transparency or secrecy. To maximize the throughput, a real-time information fee must be imposed such that the exact fraction γ_1^* of customers are informed.

Proposition 2 (Information Fee for Optimal Throughput). *Assume customers are rational. If the offered load $\rho \in [\underline{\rho}, \bar{\rho}]$, the service provider induces the optimal information level that maximizes throughput by charging an information fee $f = \bar{S}_I(\gamma_1^*)$.*

We use Figure 6(a) with $\rho = 1.1$ to illustrate Proposition 2. We first show that for the base model, if any exogenous fraction of informed customers are forced to pay the proposed information fee f, their individual utility shifts down by f units but is still nonnegative, and the utility of an uninformed customer is unchanged. Note that an informed customer's net utility reaches its minimum at $\gamma = \gamma_1^* = 0.51$, at which point the system throughput peaks (see Figure 3). Now assume that informed customers are forced to pay a fee $f = \bar{S}_I(\gamma_1^*) - \bar{S}_U(\gamma_1^*) = \bar{S}_I(\gamma_1^*)$ for inspecting the queue length. Then the utility curve of informed customers shifts down to the position displayed by the thin dotted line in Figure 6(a). As shown in the plot, informed customers still earn a nonnegative utility after paying the information fee, regardless of their fraction of the population. This implies that the additional information fee f would not alter the joining behavior of informed customers. Hence, the irrational, uninformed customers have the same equilibrium joining probability q^* as before, regardless of the proportion of informed customers. In other words, if informed customers are required to pay the fee $f = \bar{S}_I(\gamma_1^*)$, the system behaves in exactly the same way as in the base model, except that informed customers receive less utility at any information level γ.

We next show that the system with an information fee $f = \bar{S}_I(\gamma_1^*)$ reaches an equilibrium in which self-interested choices lead to only γ_1^* fraction of customers willing to pay the fee to be informed. Suppose that the system is in a state where fewer than the γ_1^* fraction of customers pay the fee and become informed, i.e., $\gamma < \gamma_1^*$. From Figure 6(a), we see that not paying the fee and staying uninformed earns zero utility and is dominated by paying the information fee and becoming informed, i.e., $\bar{S}_I(\gamma) - f > \bar{S}_U(\gamma) = 0$ for $\gamma < \gamma_1^*$. Therefore, some uninformed customers have the incentive to become informed until γ reaches γ_1^*. Next, consider the system in a state where more than γ_1^* fraction of customers are informed, i.e., $\gamma > \gamma_1^*$. In this case, paying the

information fee is not worthwhile. Saving the cost and being uninformed yields higher utility, i.e., $\bar{S}_{\mathrm{I}}(\gamma) - f < \bar{S}_{\mathrm{U}}(\gamma)$ for $\gamma > \gamma_1^*$. Therefore, some informed customers have an incentive not to pay for real-time information and become uninformed until γ reaches γ_1^*. As a result, at $\gamma = \gamma_1^*$, both options, being informed or being uninformed, are equally appealing. The system reaches an equilibrium in which the information level that maximizes the throughput is induced through customer-decentralized, rational choices under the information fee $f = \bar{S}_{\mathrm{I}}(\gamma_1^*)$.

5.2.2. Inducing the Optimal Social Welfare. According to Corollary 4, if the offered load is not high, i.e., $\rho \le \bar{\rho}$, the system attains its optimal social welfare when all customers are informed. This can be accomplished by customer self-interested choices under queue-length information transparency, since being informed is a dominant strategy. Hence, social welfare optimality can be achieved without coercion as long as the service provider reveals the real-time congestion information. By contrast, if the offered load is high, i.e., $\rho > \bar{\rho}$, the optimal social welfare is achieved when $\gamma = \gamma_0^*$, at which point uninformed customers have an incentive to become informed. The social welfare optimality thus cannot be established through decentralized decisions under information transparency. The service provider must charge an information fee to achieve the socially optimal solution.

Note that with an information fee, social welfare includes the total customer welfare and the service provider's collected fees (see Hassin and Haviv 2003, p. 49), i.e.,

$$S = \left(\sum_{i=0}^{n-1} p_i(q)\left(R - c\frac{i+1}{\mu}\right) - f\right)\cdot \gamma\Lambda + q\cdot(R - cW(q))\cdot(1-\gamma)\Lambda + f\cdot\gamma\Lambda.$$

The term $f\cdot\gamma\Lambda$ cancels out, reflecting the fact that from the perspective of the entire society, the information fee is only a transfer payment that has no effect on the value of social welfare itself but can help regulate the demand side and potentially achieve social optimality.

Proposition 3 (Information Fee for Optimal Social Welfare). *Assume customers are rational. If the offered load $\rho > \bar{\rho}$, the service provider induces the optimal information level that maximizes the social welfare by charging an information fee $f = \bar{S}_{\mathrm{I}}(\gamma_0^*)$.*

The idea behind Proposition 3 is similar to that behind Proposition 2, but with a minor difference. We use Figure 6(b), where $\rho = 2.3$ for illustration. As in the example shown in Figure 6(a), if all informed customers are required to pay the information fee (assuming they cannot choose not to pay the fee and become uninformed), their individual utility curve shifts from the bold dashed line to the thin dotted one. For any $\gamma < \gamma_0^*$, informed customers, after paying for the information, still earn positive utility, which is more than the zero utility of uninformed customers. Hence, some uninformed customers would like to inspect the queue by paying the fee. The incentive for converting from uninformed to informed vanishes until γ reaches γ_0^*, at which point being informed or uninformed receives the same individual net utility. So far, the argument for Proposition 3 is the same as that for Proposition 2. The difference comes when $\gamma > \gamma_0^*$. As illustrated by the thin dotted line in Figure 6(b), for $\gamma > \gamma_0^*$, the γ fraction of informed customers would incur negative utility after paying the information fee. This implies that under self-interested choices, paying an information fee $f = \bar{S}_{\mathrm{I}}(\gamma_0^*)$ is not individually rational for the informed customers whose fraction is more than γ_0^*. As a result, $\gamma = \gamma_0^*$ emerges as an equilibrium through customer decentralized decisions when the service provider charges an information fee $f = \bar{S}_{\mathrm{I}}(\gamma_0^*)$ for inspecting the queue.

Although our discussion focuses on achieving the optimal information level for maximizing throughput or social welfare, the service provider may have other objectives and can charge a different information fee to achieve another desired information level.

6. Effects of Heterogeneous Customer Characteristics

In the base model, we treated all customers as identical agents except for their possession of real-time congestion information. It is plausible that informed customers have other characteristics that make them different from the uninformed customers and that could also explain the difference in their information possession. For instance, informed customers might tend to own a smart phone and to be more technologically savvy and younger, and hence perhaps less patient. In this section, we explore how other heterogeneities in customer characteristics, in addition to awareness of real-time congestion, may interact with information heterogeneity to affect throughput and social welfare. Specifically, we assume that informed and uninformed customers receive a reward of, respectively, R_{I} and R_{U} from the service. Moreover, their respective unit waiting costs are c_{I} and c_{U}. In terms of joining strategy, informed customers still use a threshold policy: They join the queue if and only if its length is less than $n_{\mathrm{I}} \equiv \lfloor \nu_{\mathrm{I}} \rfloor \equiv \lfloor R_{\mathrm{I}}\mu/c_{\mathrm{I}} \rfloor$; otherwise, they balk. By contrast, uninformed customers choose their joining probability q according to the expected utility $R_{\mathrm{U}} - c_{\mathrm{U}}W(q)$. In other words, the dynamics of this extended model evolve in the same vein as the base model. As shown below, the additional customer heterogeneities do *not* change our result for system throughput.

Theorem 5 (Comparative Statics of Throughput). *Consider the model with heterogeneous customer characteristics.*

(i) *If* $0 \le q^* < 1$, *the throughput* $\lambda(q^*)$ *is strictly increasing in* γ.

(ii) *If* $q^* = 1$, *the throughput* $\lambda(q^*)$ *is strictly decreasing in* γ.

Recall that maximizing system throughput is equivalent to minimizing the idleness of the server, which results from inadequate customer arrivals and an intertemporal mismatch between capacity and demand, i.e., the mean effect and the variability effect. As we have argued, ubiquity of congestion information improves throughput by overcoming the mean effect under low offered loads by motivating uninformed customers, and by reducing the variability effect under high offered loads by better matching capacity and waiting slots with informed customers intertemporally. The influence of information prevalence on the server idleness remains robust when service rewards and unit waiting costs become heterogeneous. So, too, does our result for throughput.

By contrast, social welfare directly measures the total service rewards less the costs of delay. Thus, the result for social welfare is expected to be affected by customer heterogeneities in service evaluation and patience. The following result presents the conditions that do and do not cause a difference.

Theorem 6 (Comparative Statics of Social Welfare). *Consider the model with heterogeneous customer characteristics. Let* $\nu_I \equiv R_I\mu/c_I \ge 2$, $n_I \equiv \lfloor \nu_I \rfloor$ *and* $\nu_U \equiv R_U\mu/c_U$.

(i) *If* $q^* = 0$, *the social welfare* $S_I(q^*) + S_U(q^*)$ *is strictly decreasing in* γ.

(ii) *If* $0 < q^* < 1$, *the social welfare* $S_I(q^*) + S_U(q^*)$ *is strictly increasing in* γ *if* $\nu_I \ge \nu_U$. *Otherwise, the social welfare might be unimodal in* γ.

(iii) *If* $q^* = 1$, *the social welfare* $S_I(q^*) + S_U(q^*)$ *is strictly increasing in* γ *if* $\nu_I \ge \nu_U - ((1+\rho-\gamma\rho)/(1-\rho+\gamma\rho) - \langle\nu_I\rangle)/((1-\rho^{n_I})(1-\rho+\gamma\rho)^2/((1-\rho)\rho^{n_I})+1)$. *Otherwise, the social welfare might be unimodal.*

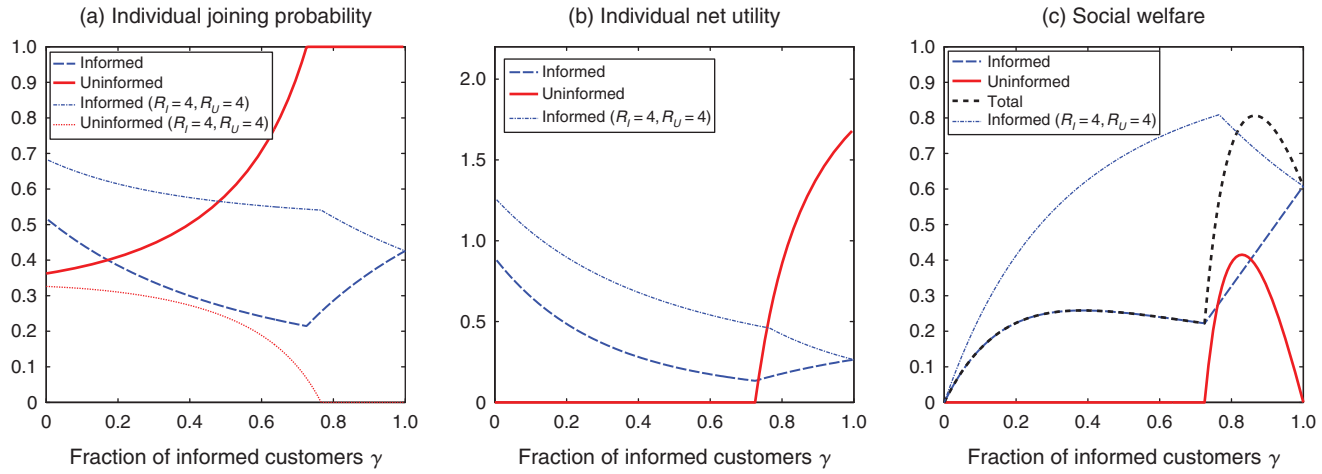
Theorem 6 reveals that if ν_I is no less than ν_U, our welfare result for the homogeneous case still holds for the heterogeneous extension. Otherwise, the social welfare can even be unimodal in the information level over the range such that $q^* \in (0,1)$ or $q^* = 1$.

The parameters ν_I and ν_U are the joining thresholds for informed and uninformed customers if they both could observe the queue length. Because customers who can tolerate a longer queue presumably have a higher valuation of service relative to their unit waiting cost, we can consider ν_I and ν_U as normalized service valuations of the two customer segments. In the homogeneous case, in which both segments of customers value the service equally, the monotonicity of the social welfare in the information level γ is primarily determined by the total utility of the informed customers. Such an effect is expected to be more salient if informed customers value the service more than the uninformed. Thus, the result in the heterogeneous case is similar to the homogeneous case if ν_I is relatively larger than ν_U as stated in Theorem 6 parts (ii) and (iii).

On the other hand, if uninformed customers value the service more after certain normalization, they desire the service more or are more patient. In either case, the monotonicity property of social welfare might be different from that in the homogeneous case. Figure 7 illustrates the differences in uninformed customers' equilibrium behavior and their effects on informed customers and social welfare. For comparison with the homogeneous case, we choose the same system parameters as those in Figure 5, except that uninformed customers receive a reward $R_U = 6$ instead of 4, whereas the reward for the informed, R_I, is still 4.

We first discuss the differences in the incentive and behavior of uninformed customers between the heterogeneous and homogeneous cases and their effect on the

Figure 7. (Color online) Example: $\mu = 1$, $R_I = 4$, $R_U = 6$, $c = 1$ and $\Lambda = 2.3$



informed customers whose joining threshold remains unchanged. For the same information level γ, uninformed customers are more enthusiastic about joining the queue in the heterogeneous case than they are in the homogeneous case because of the higher service reward.[8] In Figure 7(a), this corresponds to the fact that the solid curve representing q^* in the heterogeneous case stays higher than the dotted curve representing q^* in the homogeneous case. As a result of the increased q^*, uninformed customers compete with informed customers for waiting positions with fewer than four people ahead of them, and informed customers are less likely to join the line and earn positive utilities in the heterogeneous case than in the homogeneous case for the same information level γ. This explains why the dashed lines stay below the dash-dot lines in Figures 7(a) and 7(b), respectively. Moreover, because of the larger externalities exerted by uninformed customers (as a larger q^* at the same γ) in the heterogeneous case, the total welfare generated by the informed segment is lower than it is in the homogeneous case, as shown by the bold dashed curve and the thin dotted curve in Figure 7(c).

The difference in R_{I} and R_{U} changes the monotonicities of both customer streams' joining probability with respect to the information level. For the heterogeneous case, uninformed customers' equilibrium joining probability q^* increases in the information level γ, whereas the joining probability of informed customers first declines and then increases in γ. By contrast, for the homogeneous case, q^* decreases in the information level γ given the relatively high offered load $\rho = 2.3$. Since $\nu_{\mathrm{U}} = 6 > \nu_{\mathrm{I}} = 4$, 4 or 5 customers waiting ahead of them are welcomed by uninformed customers but is not acceptable to informed customers. As γ increases, the number of uninformed customers declines. Therefore, the fourth and fifth positions in the queue are more likely to be available, which contributes to an increasing equilibrium joining probability q^* of the uninformed customers. As q^* increases in γ, the probability that an informed customer will observe a queue shorter than n_{I} decreases. This decrease stops when q^* reaches 1, corresponding to $\gamma = 0.72$, beyond which a further increase in γ only lowers the number of uninformed customers with no way of further raising q^*. Therefore, congestion starts to be alleviated as the throughput starts to decrease (see Theorem 5(ii)). Therefore, the joining probability of informed customers increases in γ after q^* hits 1. So do the individual net utilities of informed and uninformed customers, as shown in Figure 7(b).[9]

The increasing joining probability q^* of uninformed customers in the information level γ can result in the unimodal behavior of social welfare over the range of $q^* \in (0,1)$. As γ grows, uninformed customers join the queue more often and thus inflict more negative externalities on informed customers in the heterogeneous case. By contrast, as γ increases, uninformed customers join the queue less often in the homogeneous case. Hence, in the heterogeneous case, the consumer welfare of the informed segment, which also equals the social welfare, does not necessarily increase with the growing information prevalence as it does in the homogeneous case; it may decline from a certain point, as shown in Figure 7(c).

When $q^* = 1$, individual net utility of an informed customer increases in γ and thus the total consumer welfare of the informed stream must also increase in γ. For the uninformed stream, although the net utility of an uninformed customer increases from zero, the size of the segment shrinks to zero. The total uninformed customer welfare is thus unimodal in γ. In the heterogeneous case, the welfare of the uninformed customers contributes to a larger portion of social welfare. Thus, the unimodal behavior of the uninformed customers' welfare can lead to the unimodal behavior of social welfare in γ over the range of $q^* = 1$.

In summary, when customers of different information segments exhibit different characteristics, the results for the system throughput remain the same as in the homogeneous case from our base model. However, the nonmonotonic behavior of social welfare in γ may also result from uninformed customers' high service valuation or low unit delay cost, in addition to the real-time information heterogeneity.

7. Conclusion

We have considered information heterogeneity in a service system and shown the effect of a larger proportion of informed customers in the population on various performance measures. In particular, that effect can be determined by the type of equilibrium joining behavior of uninformed customers. Perhaps surprisingly, we have shown that a larger proportion of informed customers may not necessarily benefit throughput or social welfare.

Our results suggest that the presence of uninformed customers who interact with informed customers does not necessarily jeopardize system performance. Information ignorance may not be as detrimental as one might expect. In fact, information heterogeneity helps the system in certain conditions. Our results may raise the question whether the current practice of disseminating free delay information is the most effective way to manage congestion, and whether it might be more efficient to introduce an information fee to intentionally create heterogeneity in the possession of delay information. Another implication of our results is that service providers can be better off by limiting access to real-time information about delays so as to intentionally create a mix of informed and uninformed customers. One possible way of doing that might be to target the delivery of delay announcements. For a loaded

call center, the service provider might consider making delay announcements only to a fraction of callers, e.g., loyal customers. This insight is also robust when the two customer segments have heterogeneous characteristics in service valuation and delay cost. Our findings may well justify Disney World's practice of allowing only premium customers to obtain waiting-time information about its popular attractions.[10] In that case, the amusement park may try to maximize the total satisfaction experienced by park visitors with heterogeneous real-time delay information.

Acknowledgments

The authors thank department editor Gad Allon, the anonymous associate editor, and three reviewers for their constructive and insightful suggestions, which significantly improved the paper.

Appendix. Proofs

Proof of Lemma 1. We first demonstrate a structural property of $p_i(q)$: There exists q-dependent $k \in \mathbb{N}$ such that $p_i(q)$ is decreasing in q for all $0 \le i < k$ and $p_i(q)$ is strictly increasing in q for all $i \ge k$. For $\gamma \in [0,1)$, $\lambda_U > 0$. By (5), it is clear that $p_0(q)$ is strictly decreasing in q. Moreover, since $\sum_{i=0}^{\infty} p_i(q) = 1$, there must exist some $i' \in \mathbb{N}$ such that $p_{i'}(q)$ is strictly increasing in q. Let $k = \min\{i \in \mathbb{N} \mid p_i(q) \text{ is strictly increasing in } q\}$. By the balance Equations (1) and (2), since $p_k(q)$ is strictly increasing in q, $p_{k+1}(q)$ is strictly increasing in q, and recursively, $p_i(q)$ is strictly increasing in q for all $i \ge k$. By the definition of $p_k(q)$, all $p_i(q)$'s for $0 \le i < k$ decrease in q.

Now we use the property above to show the stochastic monotonicity of Q. Fix $l \in \mathbb{N}$. If $l \le k$, $\mathbb{P}(Q(q) \ge l) = 1 - \mathbb{P}(Q(q) < l) = 1 - (\sum_{i=0}^{l-1} p_i(q))$ increases in q, because $p_i(q)$, for $i \le l-1 \le k-1$, decreases in q by part (i). If $l > k$, $\mathbb{P}(Q(q) \ge l) = \sum_{i=l}^{\infty} p_i(q)$ strictly increases in q, because $p_i(q)$, for $i \ge l > k$, strictly increases in q. The result follows by definition of the usual stochastic order in Shaked and Shanthikumar (2007). Thus, for any $0 \le q_1 < q_2 \le 1$, $Q(q_1) \le_{\text{st}} Q(q_2)$, which implies that $\mathbb{E}(Q(q_1)) \le \mathbb{E}(Q(q_2))$. By Theorem 1.A.8. in Shaked and Shanthikumar (2007), the inequality must be strict, i.e., $\mathbb{E}(Q(q_1)) < \mathbb{E}(Q(q_2))$ for $q_1 < q_2$. Since $W(q) = \mathbb{E}(Q(q))/\mu$, $W(q_1) < W(q_2)$. □

Proof of Proposition 1. The proof follows from similar arguments to Hassin and Haviv (2003, p. 46). □

Proof of Corollary 1. By Proposition 1(i), no uninformed customers join the queue if and only if $cW(0) \ge R$. Plugging in (6), we have $cW(0) \ge R \iff f(\gamma\rho) \ge \nu$, where $f(y) \equiv n + 1 + 1/(1-y) - (n+1)/(1-y^{n+1})$, $y \ge 0$. The result follows if (1) $f(y)$ is continuous and strictly increasing and (2) $f(y) = \nu$ has a unique solution. Since $\lim_{y\to 1} f(y) = n/2 + 1$, the continuity is guaranteed. The monotonicity of $f(y)$ results from the fact that

$$f'(y) = \frac{1}{(1-y)^2} - \frac{(n+1)^2 y^n}{(1-y^{n+1})^2} \ge \frac{1}{(1-y)^2} - \frac{(n+1)^2}{(1-y^{n+1})^2}\left(\frac{1}{n+1}\cdot\frac{1-y^{n+1}}{1-y}\right)^2 = 0,$$

where the last inequality is due to the inequality of arithmetic and geometric means

$$y^{n/2} = \sqrt[n+1]{1 \times y \times y^2 \times \cdots \times y^n} \le \frac{1 + y + y^2 + \cdots + y^n}{n+1} = \frac{1}{n+1}\cdot\frac{1-y^{n+1}}{1-y}.$$

Note that $f'(y) = 0$ only at a *single* point $y = 1$. Thus, $f(y)$ is strictly increasing in y. Last, since $f(0) = 1$, $\lim_{y\to\infty} f(y) = n+1$ and $\nu \in [n, n+1)$, $n \ge 1$. Therefore, $f(y) = \nu$ has a unique solution $y^*(\nu) \ge 0$. □

Proof of Corollary 2. By Proposition 1(ii) and (6), uninformed customers all join the queue if and only if

$$cW(1) \le R \iff \left(\frac{1}{1-(1-\gamma)\rho}\right)^2 - \langle\nu\rangle\left(\frac{1}{1-(1-\gamma)\rho}\right) - L(\rho,\nu) \le 0,$$

where

$$L(\rho,\nu) \equiv \frac{\langle\nu\rangle \sum_{i=0}^{n-1}\rho^i + \sum_{i=1}^{n-1}\sum_{j=0}^{i-1}\rho^j}{\rho^n} \ge 0. \tag{A.1}$$

Note that by (3),

$$0 \le \mathbb{P}(Q \ge n) = \sum_{i=n}^{\infty} p_i(q) = \sum_{i=n}^{\infty} (\rho_c(q))^n (\rho_u(q))^{i-n} p_0(q) = (\rho_c(q))^n (1-\rho_u(q))^{-1} p_0(q). \tag{A.2}$$

Then $(1-\rho_u(q))^{-1} \ge 0$ for all q. In particular, if $q = 1$, $(1-\rho_u(q=1))^{-1} = 1/(1-(1-\gamma)\rho) \ge 0$. Therefore,

$$cW(1) \le R \iff \left(\frac{1}{1-(1-\gamma)\rho}\right)^2 - \langle\nu\rangle\left(\frac{1}{1-(1-\gamma)\rho}\right) - L(\rho,\nu) \le 0 \iff 0 \le \frac{1}{1-(1-\gamma)\rho} \le \frac{\langle\nu\rangle + \sqrt{\langle\nu\rangle^2 + 4L(\rho,\nu)}}{2}.$$

The rest of the proof follows by solving for the condition on γ from the above inequality. □

Proof of Theorem 1. We first demonstrate how the two critical information levels $\gamma_0^*(\rho,\nu)$ and $\gamma_1^*(\rho,\nu)$ change with respect to the offered load ρ. Clearly, $\gamma_0^*(\rho,\nu) = y^*(\nu)/\rho$ is strictly decreasing in ρ. Since $\lim_{\rho\to\infty}\gamma_0^*(\rho,\nu) = 0$ and $\lim_{\rho\to y^*(\nu)}\gamma_0^*(\rho,\nu) = 1$, we claim $0 \le \gamma_0^*(\rho,\nu) < 1$ if and only if $\rho > y^*(\nu)$; and $\gamma_0^*(\rho,\nu) \ge 1$ if and only if $\rho \le y^*(\nu)$. On the other hand, it is easy to show that $\gamma_1^*(\rho,\nu) \ge 0 \Leftrightarrow \rho \ge 1 - 1/\nu \ge 0$ because $L(\rho,\nu) \ge 0$ and $\nu \ge 1$. Due to the same fact that $L(\rho,\nu) \ge 0$,

$$\gamma_1^*(\rho,\nu) \le 1 \iff (\langle\nu\rangle(\rho-1) + 2 - \rho)(\rho^{n+1} - 1) \ge (n+1)(\rho-1). \tag{A.3}$$

The inequality always holds if $\rho = 1$. We then only discuss the case $\rho \ne 1$. Dividing both sides by $(1-\rho)(1-\rho^{n+1}) > 0$, (A.3) can be equivalently transformed as

$$\gamma_1^*(\rho,\nu) \le 1 \iff \frac{\langle\nu\rangle(\rho-1) + 1 + 1 - \rho}{1-\rho} \le \frac{n+1}{1-\rho^{n+1}} \iff n + 1 + \frac{1}{1-\rho} - \frac{n+1}{1-\rho^{n+1}} \le \nu \iff \rho \le y^*(\nu),$$

where the last equivalence results from the fact $f(y) = n + 1 + 1/(1-y) - (n+1)/(1-y^{n+1})$ is increasing in y.

(i) Consider $0 \le \rho < 1 - 1/\nu$. In this case, $\gamma_1^*(\rho,\nu) < 0$. By Corollary 2, $q^* = 1$ for all $\gamma \in [0,1]$.

(ii) Consider $1-1/\nu \le \rho \le y^*(\nu)$. In this case, $\gamma_0^*(\rho,\nu) = y^*(\nu)/\rho \ge 1$. Thus, $q^* \ne 0$ for all $\gamma \in [0,1]$ by Corollary 1. However, $\gamma_1^*(\rho,\nu) \in [0,1]$ for $1-1/\nu \le \rho \le y^*(\nu)$, which implies that $q^* = 1$ for $\gamma_1^*(\rho,\nu) \le \gamma \le 1$. By the uniqueness of q^*, $0 < q^* < 1$ for $0 \le \gamma < \gamma_0^*(\rho,\nu)$.

(iii) Consider $\rho > y^*(\nu)$. In this case, $0 \le \gamma_0^*(\rho,\nu) < 1$. Therefore, $q^* = 0$ for $\gamma_0^*(\rho,\nu) \le \gamma \le 1$. Then, for $0 \le \gamma < \gamma_0^*(\rho,\nu)$, it is only possible that $0 < q^* \le 1$. However, by the uniqueness of q^*, it can be easily shown, by contradiction, that it must be that $0 < q^* < 1$ for $0 \le \gamma < \gamma_0^*(\rho,\nu)$. □

Proof of Lemma 2. We consider two cases separately: (i) $q^*(\gamma) = 0$ and (ii) $q^*(\gamma) \in (0,1)$.

(i) $q^*(\gamma) = 0$. By (5), $p_0(\gamma) = (1 + \sum_{i=0}^{n-1}(\gamma\rho)^{i+1})^{-1}$, which is strictly decreasing in γ.

(ii) $q^*(\gamma) \in (0,1)$. Because it is difficult to analytically solve q^* as a function of γ from $cW(q) = R$, we must verify the result indirectly. Specifically, we focus on $\rho_c = \rho[\gamma + q^*(1-\gamma)]$ instead of q^* for the monotonicity of $p_0(\gamma)$. Because we are only interested in the nontrivial cases where $\rho > 0$ and $\gamma > 0$, $\rho_c > 0$. It is sufficient to show that $d\rho_c/d\gamma > 0$ and $dp_0/d\rho_c < 0$. Then we can obtain $dp_0/d\gamma = (dp_0/d\rho_c)(d\rho_c/d\gamma) < 0$ for $q^* \in (0,1)$.

First, we verify $d\rho_c/d\gamma > 0$. When $q^* \in (0,1)$, the relationship between ρ_c and γ is determined by the equilibrium equation $cW(q^*) = (c/\mu)\sum_{i=0}^{\infty}(i+1)p_i(q^*) = R$. By (6),

$$cW(q^*) = \frac{c}{\mu}\sum_{i=0}^{\infty}(i+1)p_i(q^*) = \frac{c}{\mu}p_0(q^*)\left[\frac{1-\rho_c^n}{1-\rho_c} + \frac{\rho_c}{(1-\rho_c)^2} + \rho_c^n\left(\frac{1-n}{1-\rho_c} - \frac{1}{(1-\rho_c)^2} + \frac{1}{(1-\rho_c+\gamma\rho)^2} + \frac{n}{1-\rho_c+\gamma\rho}\right)\right] = R, \tag{A.4}$$

where $p_0(q^*) = ((1-\rho_c^n)/(1-\rho_c) + \rho_c^n/(1-\rho_c+\gamma\rho))^{-1}$. Define $\langle\nu\rangle = \nu - n$. Then Equation (A.4) is equivalent to

$$\frac{\langle\nu\rangle(\rho_c-1)\rho_c^n + \nu - \nu\rho_c + \rho_c^n - 1}{(1-\rho_c)^2\rho_c^n} = \frac{1-\langle\nu\rangle(1-\rho_c+\gamma\rho)}{(1-\rho_c+\gamma\rho)^2}, \tag{A.5}$$

where the left hand side is exactly $L(\rho_c)$ defined in Corollary 2. Recall from Equation (4) that $1-\rho_U(q^*) = 1 - q^*\lambda_c/\mu = 1-\rho_c+\gamma\rho \ge 0$. Thus, Equation (A.5) gives rise to the only positive solution to $(1-\rho_c+\gamma\rho)$, which further leads to a unique expression of $\gamma(\rho_c)$. In other words, we know from Equation (A.5) that

$$1-\rho_c+\gamma\rho = \frac{-\langle\nu\rangle + \sqrt{\langle\nu\rangle^2 + 4L(\rho_c)}}{2L(\rho_c)}$$
$$\iff \gamma(\rho_c) = \frac{1}{\rho}(2(\langle\nu\rangle + \sqrt{\langle\nu\rangle^2 + 4L(\rho_c)})^{-1} + \rho_c - 1). \tag{A.6}$$

It can be shown that $L(\rho_c)$ is strictly decreasing in ρ_c, i.e., $dL/d\rho_c < 0$ (see Lemma B1 in Online Appendix B). Therefore, $\gamma(\rho_c)$ in (A.6) is strictly increasing in ρ_c, i.e., $d\gamma/d\rho_c > 0$. Moreover, since the inverse function of a strictly increasing function is also strictly increasing, $d\rho_c/d\gamma > 0$.

Second, we show $dp_0/d\rho_c < 0$. We write $p_0(q^*)$ as a function of ρ_c:

$$p_0(q^*) \overset{(5)}{=} \left(\frac{1-\rho_c^n}{1-\rho_c} + \frac{\rho_c^n}{1-\rho_c+\gamma\rho}\right)^{-1} \overset{(A.6)}{=} \left[\sum_{i=0}^{n-1}\rho_c^i + \frac{\rho_c^n}{2}\langle\nu\rangle + \sqrt{\rho_c^{2n}\langle\nu\rangle^2/4 + \rho_c^{2n}L(\rho_c)}\right]^{-1}.$$

Recall that $\rho_c > 0$. Hence, if $\rho_c^{2n}L(\rho_c)$ is strictly increasing in ρ_c, we can easily show that $p_0(q^*)$ is strictly decreasing in ρ_c. Note

$$\rho_c^{2n}L(\rho_c) = \rho_c^n\frac{\langle\nu\rangle(1-\rho_c)(1-\rho_c^n) + (1-\rho_c)\sum_{i=0}^{n-1}(1-\rho_c^i)}{(1-\rho_c)^2} = \rho_c^n\left(\langle\nu\rangle\sum_{i=0}^{n-1}\rho_c^i + \sum_{i=1}^{n-1}\sum_{j=0}^{i-1}\rho_c^j\right), \tag{A.7}$$

which confirms that $\rho_c^{2n}L(\rho_c)$ is indeed strictly increasing in $\rho_c > 0$ and implies that $dp_0/d\rho_c < 0$. □

Proof of Theorem 2. (i) Recall that $\lambda(q) = \mu(1-p_0(q))$. Hence, for $q^* \in [0,1)$, it is obvious that $\lambda(q^*)$ strictly increases in γ, since we have shown that $p_0(q^*)$ strictly decreases in γ in Lemma 2.

(ii) If $q^* = 1$, $\rho_c(q^*=1) = \rho$ and $\rho_U(q^*=1) = (1-\gamma)\rho$. Then

$$\lambda(q^*=1) \overset{(7)}{=} \left(\sum_{i=0}^{n-1}p_i(1)\right)\lambda_I + \lambda_U \overset{(3),(5)}{=} \left(\frac{1-\rho^n}{1-\rho} + \frac{\rho^n}{1-(1-\gamma)\rho}\right)^{-1} \cdot \left(\frac{1-\rho^n}{1-\rho}\right)\gamma\Lambda + (1-\gamma)\Lambda = \frac{-1+\rho-\gamma\rho+\gamma\rho^n}{-1+\rho-\gamma\rho+\gamma\rho^{n+1}}\Lambda.$$

Differentiating $\lambda(q^*=1)$ w.r.t. γ, we have $(d/d\gamma)\lambda(q^*=1) = -(\rho^n(1-\rho)^2/(-1+\rho-\gamma\rho+\gamma\rho^{n+1})^2)\Lambda < 0$ for $\rho > 0$. Thus, $\lambda(q^*=1)$ is strictly decreasing in γ when $q^* = 1$. □

Proof of Corollary 3. The result immediately follows by combining Theorems 1 and 2. □

Proof of Theorem 3. (i) We first verify that $d\bar{S}_I(q^*(\gamma'))/d\gamma < 0$ for any given γ' such that $0 \le q^*(\gamma') < 1$. Lemma B2 in Online Appendix B states that for any such γ' there exists $k \le n-1$ such that $dp_i(q^*(\gamma'))/d\gamma < 0$ for $0 \le i \le k$ and $dp_i(q^*(\gamma'))/d\gamma \ge 0$ for $k < i < n$. Therefore,

$$\frac{d\bar{S}_I(q^*(\gamma'))}{d\gamma} \le \sum_{i=0}^{k}\frac{dp_i(q^*(\gamma'))}{d\gamma}\left(R - c\frac{k+1}{\mu}\right) + \sum_{i=k+1}^{n-1}\frac{dp_i(q^*(\gamma'))}{d\gamma}\left(R - c\frac{k+1}{\mu}\right) = \left(R - c\frac{k+1}{\mu}\right)\sum_{i=0}^{n-1}\frac{dp_i(q^*(\gamma'))}{d\gamma} < 0,$$

where the last inequality results from Proposition B1(i) (see Online Appendix B).

When $q^* = 1$, $\bar{S}_I(q^*=1) = \sum_{i=0}^{n-1}p_i(q^*=1)(R - c((i+1)/\mu)) = \sum_{i=0}^{n-1}\rho^i p_0(q^*=1)(R - c((i+1)/\mu))$. Since $p_0(q^*=1) = ((1-\rho^n)/(1-\rho) + \rho^n/(1-(1-\gamma)\rho))^{-1}$ strictly increases in γ, so does $\bar{S}_I(q^*=1)$.

(ii) By definition, $\bar{S}_U(q^*) = 0$ if $0 \le q^* < 1$. Thus, we need only consider the case where $q^* = 1$,

$$\bar{S}_U(q^*=1) = \sum_{i=0}^{\infty}p_i(q^*=1)\left(R - c\frac{i+1}{\mu}\right) = p_0(q^*=1)\left(\sum_{i=0}^{n-1}\rho^i\left(R - c\frac{i+1}{\mu}\right) + \sum_{i=n}^{\infty}\rho^n((1-\gamma)\rho)^{i-n}\left(R - c\frac{i+1}{\mu}\right)\right).$$

First, $p_0(q^*=1) = ((1-\rho^n)/(1-\rho) + \rho^n/(1-\rho+\gamma\rho))^{-1}$ strictly increases in γ. Second, $R - c((i+1)/\mu) < 0$ for $i \ge n$, which implies $\rho^n((1-\gamma)\rho)^{i-n}(R - c((i+1)/\mu))$ increases in γ. Therefore, $\bar{S}_U(q^*=1)$ strictly increases in γ. □

Proof of Theorem 4. We prove the results case by case for $q^*=0$, $q^*\in(0,1)$ and $q^*=1$, respectively. By definition, $S_U(q^*)=0$ for all γ if $q^*=0$. By Proposition 1(iii), $q^*\in(0,1)$ must satisfy that $R=cW(q^*)$, hence, $S_U(q^*)=0$ for all γ if $q^*\in(0,1)$. Then, $S_U(q^*)>0$ can only happen when $q^*=1$.

(a) When $q^*=0$, we have $S_I(q^*=0)=(\sum_{i=0}^{n-1}p_i(0)(R-c\cdot(i+1)/\mu))\cdot\gamma\Lambda=(g(\gamma\rho)+\langle\nu\rangle)c$, where

$$g(z)\equiv\frac{1}{z-1}-\frac{n+1}{z^{n+1}-1}-\frac{vz-v}{z^{n+1}-1}.$$

If $n=1$, $S_I(q^*=0)=(-\langle\nu\rangle/(1+\gamma\rho)+\langle\nu\rangle)c$, which strictly increases in γ. Hence so does S_I+S_U for $1\le\nu<2$.

We next prove that $g(z)$ is a decreasing function of z for $z\ge y^*(\nu)$ when (i) $n=2$, 3; and (ii) $n\ge4$. The proof relies on an important property of $y^*(\nu)$ as shown by Lemma B3 in Online Appendix B. That is, for $\nu\ge2$ and $\bar\nu=\nu+i$ for any $i\in N$, it must be that $y^*(\bar\nu)>y^*(\nu)\ge1$.

(i) When $n=2\Leftrightarrow\nu\in[2,3)$, we have $g(z)=(z-\langle\nu\rangle)/(1+z+z^2)$, which is a decreasing function if

$$g'(z)=\frac{-z^2+2\langle\nu\rangle z+\langle\nu\rangle+1}{(z^2+z+1)^2}<0\ \Leftrightarrow\ z>\langle\nu\rangle+\sqrt{\langle\nu\rangle^2+\langle\nu\rangle+1}.$$

Furthermore, we can analytically solve $y^*(\nu)$ defined in Corollary 1 as $y^*(\nu)=(\langle\nu\rangle+\sqrt{4-3\langle\nu\rangle^2})/(2(1-\langle\nu\rangle))$. By basic algebra, we can prove that $(\langle\nu\rangle+\sqrt{4-3\langle\nu\rangle^2})/(2(1-\langle\nu\rangle))\ge\langle\nu\rangle+\sqrt{\langle\nu\rangle^2+\langle\nu\rangle+1}$ for $\langle\nu\rangle\in[0,1)$.

When $n=3\Leftrightarrow\nu\in[3,4)$, we have

$$\begin{aligned}g(z)&=\frac{z+z^2}{1+z+z^2+z^3}+\frac{z-\langle\nu\rangle}{1+z+z^2+z^3}\\&=\frac{1}{1/z+z}+\frac{z-\langle\nu\rangle}{1+z+z^2}\frac{1}{1+1/(z^{-3}+z^{-2}+z^{-1})},\end{aligned}$$

where $1/(1/z+z)$ and $1/(1+1/(z^{-3}+z^{-2}+z^{-1}))$ are clearly positive and decreasing functions of $z\ge y^*(\nu)\ge1$; from the result for the case of $n=2$ and Lemma B3, we have $(z-\langle\nu\rangle)/(1+z+z^2)$ is a positive and decreasing function for $z\ge y^*(\nu)\ge y^*(\nu-1)\ge1>\langle\nu\rangle$.

Thus, $g(z)$ is strictly decreasing for $z\ge y^*(\nu)$ when $n=2$ or 3.

(ii) When $n\ge4$, the first-order derivative of $g(z)$ is

$$\begin{aligned}g'(z)&=-\frac{1}{(z-1)^2}+\frac{(n+1)^2z^n}{(z^{n+1}-1)^2}+\frac{\nu(z-1)(nz^n-\sum_{i=0}^{n-1}z^i)}{(z^{n+1}-1)^2}\\&\overset{\nu<n+1}{<}\frac{1}{(z^{n+1}-1)^2}\left((nz^{n+1}+1)(n+1)-\frac{(z^{n+1}-1)^2}{(z-1)^2}\right).\end{aligned}$$

Then we just need to prove $(nz^{n+1}+1)(n+1)\le(z^{n+1}-1)^2/(z-1)^2$. Clearly, when $z=1$, we have $(nz^{n+1}+1)(n+1)=(z^{n+1}-1)^2/(z-1)^2$. To prove $(nz^{n+1}+1)(n+1)\le(z^{n+1}-1)^2/(z-1)^2$ for $z\ge y^*(\nu)\ge1$, we need only prove $d((nz^{n+1}+1)\cdot(n+1))/dz\le d((z^{n+1}-1)^2/(z-1)^2)/dz$ for $z\ge1$.

$$\begin{aligned}&\frac{d((z^{n+1}-1)^2/(z-1)^2)}{dz}\\&=2\frac{(z^{n+1}-1)(nz^{n+1}-nz^n-z^n+1)}{(z-1)^3}=2\sum_{i=0}^{n}z^i\left(\sum_{i=0}^{n-1}\sum_{j=i}^{n-1}z^j\right)\\&=2\frac{z^n}{z^{1/2}}\frac{\sum_{i=0}^{n}z^i}{z^{n/2}}\left(\frac{\sum_{j=0}^{n-1}z^j}{z^{(n-1)/2}}+\frac{\sum_{j=1}^{n-1}z^j}{z^{n/2}}\frac{z^{n/2}}{z^{(n-1)/2}}\right.\\&\quad\left.+\frac{\sum_{j=2}^{n-1}z^j}{z^{(n+1)/2}}\frac{z^{(n+1)/2}}{z^{(n-1)/2}}+\cdots+\frac{\sum_{j=n-1}^{n-1}z^j}{z^{n-1}}\frac{z^{n-1}}{z^{(n-1)/2}}\right)\\&\ge2\frac{z^n}{z^{1/2}}(n+1)(n+(n-1)z^{\frac12}+(n-2)z+\cdots+z^{(n-1)/2})\\&\qquad\text{(using } z^k+z^{-k}\ge2)\\&=2z^n(n+1)\sum_{i=1}^{n}(z^{-1/2}+1+z^{1/2}+z\cdots+z^{(i-2)/2})\\&=2z^n(n+1)((2z^{-1/2}+2z^{1/2}+3+2z^{-1/2}+z)\\&\quad+\sum_{i=5}^{n}(z^{-1/2}+1+z^{1/2}+z+\cdots+z^{(i-2)/2}))\\&\ge2z^n(n+1)\left(10+\sum_{i=5}^{n}i\right)\\&\qquad\text{(using } 2z^{-1/2}+z\ge3 \text{ and } z^{1/2}+z^{-1/2}\ge2 \text{ when } z\ge1)\\&=n(n+1)^2z^n\\&=\frac{d((nz^{n+1}+1)(n+1))}{dz}.\end{aligned}$$

Note that the last inequality and the second to last equality hold for $n=4$, if we define $\sum_{i=5}^{4}i\equiv0$.

(b) When $q^*\in(0,1)$, again let $\rho_c\equiv\rho(\gamma+q^*(1-\gamma))$. Then, we have

$$\begin{aligned}S_I(q^*)&=\bar S_I(q)\cdot\gamma\Lambda\\&=\frac{c\Lambda}{\mu}p_0(q^*)\left(\nu\frac{1-\rho_c^n}{1-\rho_c}-\frac{1-(n+1)\rho_c^n+n\rho_c^{n+1}}{(1-\rho_c)^2}\right)\gamma(\rho_c)\\&\overset{(A.5)}{=}c\rho p_0(q^*)\rho_c^nL(\rho_c)\gamma(\rho_c).\end{aligned}\tag{A.8}$$

Substitute $p_0(q^*)$ and $\gamma(\rho_c)$ with (5) and (A.6), respectively. After some algebraic manipulations,

$$\begin{aligned}&S_I(q^*)\\&=c\left(1-\left(\nu(\langle\nu\rangle+\sqrt{\langle\nu\rangle^2+4L(\rho_c)})\right)\right.\\&\quad\left.\cdot\left(2\rho_c^nL(\rho_c)+\left(\langle\nu\rangle\rho_c^n+\frac{1-\rho_c^n}{1-\rho_c}\right)(\langle\nu\rangle+\sqrt{\langle\nu\rangle^2+4L(\rho_c)})\right)^{-1}\right)\end{aligned}\tag{A.9}$$

$$\begin{aligned}&=c\left(1-\frac{\nu}{(1-\rho_c^n)/(1-\rho_c)+(\rho_c^n/2)(\langle\nu\rangle+\sqrt{\langle\nu\rangle^2+4L(\rho_c)})}\right)\\&=c(1-\nu p_0(q^*)),\end{aligned}\tag{A.10}$$

where the last equality is due to the fact that

$$\begin{aligned}p_0(q^*)&\overset{(5)}{=}\left(\frac{1-\rho_c^n}{1-\rho_c}+\frac{\rho_c^n}{1-\rho_c+\gamma(\rho_c)\rho}\right)^{-1}\\&\overset{(A.6)}{=}\left(\frac{1-\rho_c^n}{1-\rho_c}+\frac{\rho_c^n}{2}(\langle\nu\rangle+\sqrt{\langle\nu\rangle^2+4L(\rho_c)})\right)^{-1}.\end{aligned}$$

Recall that $p_0(q^*)$ strictly decreases in γ as shown in Lemma 2. Then, (A.10) indicates that $S_I(q^*)$ strictly increases in γ.

(c) Lastly, when $q^*=1$, $\rho_c=\rho$. Then, we have

$$\begin{aligned}&S_I(q^*=1)+S_U(q^*=1)\\&=\left(\sum_{i=0}^{n-1}p_i(1)\left(R-c\frac{i+1}{\mu}\right)\right)\cdot\gamma\Lambda\\&\quad+\left(\sum_{i=0}^{\infty}p_i(1)\left(R-c\frac{i+1}{\mu}\right)\right)\cdot(1-\gamma)\Lambda\\&=\frac{c\Lambda}{\mu}p_0(1)\left[\left(\frac{n(1-\rho)-(1-\rho^n)}{(1-\rho)^2}+\langle\nu\rangle\frac{1-\rho^n}{1-\rho}\right)\right.\\&\quad\left.+\rho^n\left(\frac{\gamma-1}{(1-\rho+\gamma\rho)^2}+\langle\nu\rangle\frac{1-\gamma}{1-\rho+\gamma\rho}\right)\right].\end{aligned}$$

First, $p_0(q^*=1)=((1-\rho^n)/(1-\rho)+\rho^n/(1-(1-\gamma)\rho))^{-1}$ strictly increases in γ. Second, because $S_{\mathrm{I}}(q^*=1)+S_{\mathrm{U}}(q^*=1)>0$, the term in the square bracket is positive. Moreover, it also strictly increases in γ, since

$$\frac{\partial}{\partial\gamma}\left(\frac{\gamma-1}{(1-\rho+\gamma\rho)^2}+\langle\nu\rangle\frac{1-\gamma}{1-\rho+\gamma\rho}\right)=\frac{1-\langle\nu\rangle+(1+\langle\nu\rangle)(1-\gamma)\rho}{(1-\rho+\gamma\rho)^3}\geq 0,$$

where the last inequality results from the fact that $1-\rho+\gamma\rho>0$ implied by (A.2) and $\langle\nu\rangle\in[0,1)$. As a result, $S_{\mathrm{I}}(q^*=1)+S_{\mathrm{U}}(q^*=1)$ is strictly increasing in γ. □

Proof of Corollary 4. The result immediately follows by combining Theorems 1 and 4. □

Proof of Proposition 2. Customers evaluate their options, between being informed after paying a fee f and staying uninformed, and then pick the one that maximizes their net utility. Note that $\bar{S}_{\mathrm{I}}(\gamma_1^*)-f=\bar{S}_{\mathrm{U}}(\gamma_1^*)=0$. Therefore, no one would have an incentive to deviate at $\gamma=\gamma_1^*$ and thus $\gamma=\gamma_1^*$ is an equilibrium. We will show that at any information level $\gamma\neq\gamma_1^*$, the informed customers or the uninformed customers have an incentive to deviate.

Assume that $0\leq\gamma<\gamma_1^*$. Recall that $\rho\in[\underline{\rho},\bar{\rho}]$, $0<q^*(\gamma)<1$ and then $\bar{S}_{\mathrm{U}}(\gamma_1^*)=0$ for $0<\gamma<\gamma_1^*$. By Theorem 3, $0=\bar{S}_{\mathrm{I}}(\gamma_1^*)-f<\bar{S}_{\mathrm{I}}(\gamma)-f$, with the latter decreasing in γ. Thus, $\bar{S}_{\mathrm{U}}(\gamma)<\bar{S}_{\mathrm{I}}(\gamma)-f$ for $0<\gamma<\gamma_1^*$. Then, uninformed customers would have an incentive to deviate and pay the information access fee f to become informed.

If $\gamma_1^*<\gamma\leq 1$, it is the informed customers who want to deviate. To see this, we show that $\bar{S}_{\mathrm{I}}(\gamma)-f<\bar{S}_{\mathrm{U}}(\gamma)$ for $\gamma_1^*<\gamma\leq 1$. Note that in this range of γ, $q^*=1$. We have

$$\begin{aligned}&\bar{S}_{\mathrm{I}}(\gamma)-f-\bar{S}_{\mathrm{U}}(\gamma)\\&=\sum_{i=0}^{n-1}p_i(q^*=1)\left(R-c\frac{i+1}{\mu}\right)-f-\sum_{i=0}^{\infty}p_i(q^*=1)\left(R-c\frac{i+1}{\mu}\right)\\&=\frac{c}{\mu}\cdot\frac{1-(v-n)(1-\rho+\gamma\rho)}{(1-\rho+\gamma\rho)(1+\gamma\rho((\rho^n-1)/(\rho-1)))}\rho^n-f.\end{aligned}$$

Since f is a constant, $\bar{S}_{\mathrm{I}}(\gamma)-f-\bar{S}_{\mathrm{U}}(\gamma)$ apparently decreases in γ. Thus, for $\gamma_1^*<\gamma\leq 1$, $\bar{S}_{\mathrm{I}}(\gamma)-f-\bar{S}_{\mathrm{U}}(\gamma)<\bar{S}_{\mathrm{I}}(\gamma_1^*)-f-\bar{S}_{\mathrm{U}}(\gamma_1^*)=0\iff\bar{S}_{\mathrm{I}}(\gamma)-f<\bar{S}_{\mathrm{U}}(\gamma)$. □

Proof of Proposition 3. We omit the proof since it follows the same idea as Proposition 2. □

Proof of Theorems 5 and 6. The proofs of these two theorems are similar to those of Theorems 2 and 4, but are more involved. See Online Appendix A for the details. □

Endnotes

[1] The poll results are available at http://www.gasbuddy.com/GB_Past_Polls.aspx?poll_id=720 (last accessed May 12, 2017).

[2] Although a customer needs to buy a day pass, there is no additional charge for any ride or attraction. Therefore, the entrance price is not likely to be a factor when a customer chooses a ride.

[3] If the uninformed customers are unaware of the information level γ, it is difficult to disentangle the effects of delay information heterogeneity from those of the unawareness of γ.

[4] Even if we consider regular customer's traveling cost, all of our results still hold as shown in Theorems 5 and 6. In this case, let $R_{\mathrm{I}}=R$, $R_{\mathrm{U}}=R-t$ and $c_{\mathrm{I}}=c_{\mathrm{U}}=c$, where t is a fixed traveling cost. Hence, $\nu_{\mathrm{I}}\geq\nu_{\mathrm{U}}$. By Theorems 5 and 6, results in the base model still hold for this case when $\nu_{\mathrm{I}}\geq\nu_{\mathrm{U}}$.

[5] The cut-off point $\gamma_0^*(\rho,\nu)$ can be shown to be always nonnegative but can be larger or smaller than 1. If $\gamma_0^*(\rho,\nu)>1$, the case $q^*=0$ is moot, i.e., there exists no $\gamma\in[0,1]$ such that $q^*=0$.

[6] The cut-off point $\gamma_1^*(\rho,\nu)$ can be negative. If $\gamma_1^*(\rho,\nu)<0$, then the case $q^*=1$ holds for all $\gamma\in[0,1]$. Moreover, $\gamma_1^*(\rho,\nu)$ can be larger than 1. If $\gamma_1^*(\rho,\nu)>1$, the case $q^*=1$ is moot, i.e., there exists no $\gamma\in[0,1]$ such that $q^*=1$.

[7] If the provider aims to induce the optimal social welfare for $\rho>y^*(\nu)$, $cW(q^*(\gamma^*)=0)=R$ at $\gamma^*=\gamma_0^*$. Thus, uninformed customers are actually indifferent between joining or balking if they are told that the expected delay is $W(q^*(\gamma^*))$. To ensure that no uninformed individuals join, the provider can announce a slightly longer delay than $W(q^*(\gamma^*))$ in practice.

[8] In fact, it can be rigorously proved that if $\nu_{\mathrm{U}}\geq\lfloor\nu_{\mathrm{I}}\rfloor+1$, uninformed customers never choose to always balk in equilibrium regardless of the offered load $\rho=\Lambda/\mu$. That is not the case if $\nu_{\mathrm{U}}=\nu_{\mathrm{I}}$.

[9] Note that since $R_{\mathrm{U}}>R_{\mathrm{I}}$, an uninformed customer can receive higher individual utility than an informed customer when $q^*=1$, as shown in Figure 7(b). This situation never occurs in the homogeneous case.

[10] A TouringPlans.com Premium Subscription is needed for access to the data on waiting times and crowds through a mobile app.

References

Aflaki A, Feldman P, Swinney R (2015) Should consumers be strategic? Working paper, Fuqua School of Business, Duke University, Durham, NC.

Allon G, Bassamboo A (2011) The impact of delaying the delay announcements. *Oper. Res.* 59(5):1198–1210.

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.

Debo L, Veeraraghavan S (2014) Equilibrium in queues under unknown service times and service value. *Oper. Res.* 62(1):38–57.

Edelson NM, Hilderbrand DK (1975) Congestion tolls for poisson queuing processes. *Econometrica* 43(1):81–92.

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.

Hassin R (1986) Consumer information in markets with random product quality: The case of queues and balking. *Econometrica* 54(5):1185–1195.

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

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

Huang T, Chen Y-J (2015) Service systems with experience-based anecdotal reasoning customers. *Prod. Oper. Management* 24(5):778–790.

Huang T, Allon G, Bassamboo A (2013) Bounded rationality in service systems. *Manufacturing Service Oper. Management* 15(2): 263–279.

Kremer M, Debo L (2016) Inferring quality from wait time. *Management Sci.* 62(10):3023–3038.

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

Plambeck EL, Wang Q (2013) Implications of hyperbolic discounting for optimal pricing and scheduling of unpleasant services that generate future benefits. *Management Sci.* 59(8): 1927–1946.

Shaked M, Shanthikumar JG (2007) *Stochastic Orders*. Springer Series in Statistics (Springer, New York).

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

Veeraraghavan S, Debo LG (2011) Herding in queues with waiting costs: Rationality and regret. *Manufacturing Service Oper. Management* 13(3):329–346.

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