ISSN 1866-3494

September 24, 2018

ESMT Working Paper 18-04

Opaque queues: Service systems with rationally inattentive customers

Caner Canyakmaz, ESMT Berlin

Tamer Boyaci, ESMT Berlin

Opaque Queues: Service Systems with Rationally Inattentive Customers

Caner Canyakmaz
ESMT Berlin, European School of Management and Technology, caner.canyakmaz@esmt.org,

Tamer Boyacı
ESMT Berlin, European School of Management and Technology, tamer.boyaci@esmt.org,

Classical models of service systems with rational and strategic customers assume queues to be either fully visible or invisible. In practice, however, most queues are only "partially visible" or "opaque", in the sense that customers are not able to discern precise queue length upon arrival. This is because assessing queue length and associated delays require time, attention, and cognitive capacity which are all limited. Service firms may influence this information acquisition process through their choices of physical infrastructure and technology. In this paper, we study rational queueing behavior when customers have limited time and attention. Following the theory of rational inattention, customers optimally select the type and amount of information to acquire and ignore any information that is not worth obtaining, trading off the benefits of information against its costs before deciding to join. We establish the existence and uniqueness of a customer equilibrium and delineate the impact of information costs. We show that although limited attention is advantageous for a firm in a congested system that customers value highly, it can be detrimental for less popular services that customers deem unrewarding. These insights remain valid when the firm optimally selects the price. We also discuss social welfare implications and provide prescriptive insights regarding information provision. Our framework naturally bridges visible and invisible queues, and can be extended to analyze richer customer behavior and complex queue structures, rendering it a valuable tool for service design.

Key words: service operations, rational inattention, strategic customers, rational queueing, information costs, system throughput, social welfare

1. Introduction

In many service settings, customers do not have perfect information about the queue size and associated delays. This stems mainly from the potential *information frictions* that are present, due to the physical environment and/or the limited cognitive capabilities of the customers. In some instances, like in supermarkets and ticketing booths for events, even though the queue is in theory completely observable, physical obstructions such as shelves, walls or pillars may make it difficult or impossible to judge the extent of the queue. Such designs could even be due to strategic firm decisions. For many other services like call-centers or health services, information about the queue size may not be readily available. Today, advances in information technology enable service providers or third-party firms to provide real time delay information or predictions of it. For example, many hospitals in Canada and US post emergency room (ER) wait times online. In a similar vein, Disney offers a mobile app that provides wait time information for rides in its theme parks. Third-party firms like Touring Plans utilize advanced analytics techniques and real time customer reporting to provide similar estimates. Even then, the provided information does not completely resolve customer's uncertainty about queue size and delays. First, the delay information may be deemed as not necessarily accurate, prompting customers to privately learn and validate it. For example, in the context of ER wait times, Ang et al. (2015) empirically show that hospital posted wait times are extremely unreliable, and can be off by as much as an hour and a half for much of the time. Second, and perhaps more importantly, customers have limited time and attention to devote to acquiring such data and processing the obtained data in useful information. Since information acquisition and processing is "costly", customers are bound to make decisions based on partial information. Our paper is based on this premise of customer behavior in services.

In classical models of queuing behavior, customers are assumed to be rational and strategic. That is, they maximize expected utility, which is negatively influenced by expected delays. They are also strategic, implying that they consider the actions of other customers when deciding to join or to balk. The equilibrium analysis of these models mainly differs depending on queue visibility, termed as *visible* (observable) and *invisible* (unobservable) queues. The visible and invisible queues are canonical and orthogonal representations. Visible queues captures the case where customers can freely and accurately assess the length of the queue upon arrival and hence can also estimate the expected waiting costs perfectly. Invisible queues, on the other hand, represents the other extreme scenario under which customers cannot observe the length of the queue and hence cannot estimate waiting costs. In this case, customers act completely based on beliefs. However, as emphasized above, in practice most queues are arguably *opaque*, i.e., they are only partially visible to the customers. In the end, determining the exact number of customers ahead in the queue requires time, attention, and cognitive capacity, which are all limited in quantity. Rational customers should trade-off the benefits of acquiring better information with the cost associated with it. It may simply turn out to be impossible to make these calculations with certainty and in some cases it may not make sense to spend any effort to determine it.

In this paper, we develop a general framework for *opaque* queues. We capture the salient characteristics of limited attention and cognitive capabilities of customers through a model based on the rapidly growing theory of *rational inattention*[1] in economics. Following the seminal works of Sims (2003) and Sims (2006), this theory quantifies information as reduction in Shannon entropy and assumes that utility-maximizing customers optimally select the *type* and *quantity* of information they need, and ignore the information that is not worth obtaining. In other words, information acquisition process is completely endogenized. Rationally inattentive customers know that they are not going to be able to resolve all uncertainties and make perfect queuing decisions, but they are able to decide (optimally) on what to learn and to what detail. Naturally, this selection depends on all key factors - how much time and attention customers have (i.e., information costs), prior beliefs, as well the nature of uncertainties faced (i.e., what's at stake). We embed rationally inattentive behavior of customers in a strategic queuing model that naturally bridges the well-studied visible and invisible queues. Utilizing this framework of opaque queues, we seek answers to the following fundamental research questions:

- What equilibrium behavior will prevail (if any) if customers have limited attention and optimally acquire costly information about queue size prior to joining or balking? How is the equilibrium shaped by service characteristics such as service rates, delay costs, rewards as well as information costs?
- How does limited attention and information costs impact throughput, social welfare as well as the firm's optimal pricing decision?
- Can the firm benefit from customer's limited attention? When does it have the most detrimental effects? What are the implications on the firm's queue information provision strategy?

In order to address these questions, we develop a baseline model of a single service provider facing a homogeneous population of customers with limited attention. Customers arriving to the queue are not able to discern the exact number of customers ahead of them and hence the expected delay cost. They

[1] Throughout the paper we use limited attention and rational inattention interchangeably.

have some belief about the queue size (which is formed in equilibrium) and can improve their assessment by optimally acquiring and processing costly information, and decide to join or balk accordingly. We show that there is a unique equilibrium that emerges and establish the directional properties with respect to service characteristics. We find that impact of information costs is more involved, resulting in non-trivial and possibly non-monotone queuing behavior in equilibrium. Elaborating more on such cases, we identify service characteristics that make opaque queues particularly preferable for the service provider. At the same time, we discuss optimal pricing and welfare implications of limited attention. Finally, we demonstrate the versatility of our proposed framework by considering natural extensions of the baseline model including finite queue capacity and uncertainty in other service characteristics such as service reward and service rates.

Our contribution to the literature is three-fold. First, from a theoretical perspective, we develop a micro-founded, tractable framework for service systems that accounts for customers limited attention and information processing capabilities. With visible and invisible queues sitting on the two limiting ends, this framework covers the entire spectrum of “queue transparency” and provides a unified view of the effects of information costs on service performance, firm profitability and social welfare. It offers a natural and systematic way to describe customer’s limitations and behavioral adjustments to changes in the service environment. Furthermore, there are natural connections with approaches based on search costs and bounded rationality, which provides a rich context to interpret and position our results.

Second, utilizing our framework, we provide descriptive results on rational customers’ queuing decisions in the presence of information frictions. On one hand, these results confirm the validity of intuitive facts such as to why a customer who expects a higher utility (due to higher service reward, lower price, lower delay costs) or who encounters a faster queue (due to faster service rate) is more likely to join the service, regardless of how constrained she is in attention. On the other hand, we provide a normative foundation for more subtle and less intuitive findings, especially related to the impact of information costs. In particular, the effects of information costs on equilibrium joining rates (i.e. throughput) can be non-monotone and rather complex. Nevertheless, we are able to glean structural observations and identify clearly the conditions under which throughput is unimodal in the information costs.

Our descriptive results have prescriptive insights for practicing managers. We find that service firms should be most concerned about information costs when the reward that customers obtain from joining the service is positively correlated with congestion (demand or offered load). In particular, when customers attach high value to the service that is also popular (high demand), the firm can benefit from the limited attention of customers. The throughput of an opaque queue is higher than fully visible or invisible cases. In sharp contrast, when the service firm faces relatively low demand from customers who do not particularly value the service, limited attention of customers can be detrimental. In this case, the firm is better off by either making the queue fully visible or by completely obstructing it. Interestingly, these insights remain valid when the service firm sets prices optimally. As a matter of fact, customer inattentiveness has accentuated effects in this case. This is because the firm can benefit from inattention by overcharging customers when they are more willing to join the queue. Pricing also enables the firm to moderate the throughput losses when customers are less willing to join the queue. From a welfare perspective, consumer surplus suffers from limited attention. However, when firm surplus is taken into account, welfare implications can change. Obstructing queue information partially can result in win-win outcomes for both the consumer and the firm (and hence improve social welfare) when firm profitability (margin), congestion, and customer's reward from service are all high.

Finally, we believe our approach and proposed framework can serve as the building block for modeling and analyzing richer contexts that involve strategic customer behavior in service industries. We present a number of extensions in this direction. For example, limited attention and costly information acquisition does not need to be confined to queue size or delays. Customer can spend time and effort to learn additional factors that may be uncertain, such as service speed or customer reward (service quality). In this case, customers have to allocate their attention among different elements and acquire information appropriately before making their decisions. It may even be easier to learn some elements (e.g., queue size) than others (e.g., speed). In a similar vein, the service system can have physical capacity constraints, the customer behavior may be more complex involving potential retrials. These can also be incorporated into our proposed framework, which would pave the way for new applications that deepen our understanding on how consumer's limited attention and information costs impact service design and performance.

2. Literature Review

There is a long standing literature on strategic behavior in queueing, starting with the pioneering works of Naor (1969) and Edelson and Hilderbrand (1975) for the cases of visible and invisible queues respectively. Hassin and Haviv (2003) and Hassin (2016) present excellent coverage of various extensions and comprehensive review of related literature. There are two streams within this literature that our work relates to the most: i) bounded rationality and ii) information acquisition and control.

Bounded rationality postulates that customers are not always able to make perfect rational choices and therefore make errors. Huang et al. (2013) adapts this to a service context by assuming that customers are not able to perfectly estimate their expected waiting times and investigate the revenue, pricing and welfare implications for both visible and invisible queues. As customary in the bounded rationality literature, customers face additive noise terms in their estimation and make joining decisions according to the multinomial logit (MNL) formula. The degree of bounded rationality is an exogenous parameter (standard deviation of the additive noise term) in this model. On one extreme, customers are fully rational, recovering the Naor (1969) and Edelson and Hilderbrand (1975) models. On the other extreme, customers are fully irrational and join or balk with equal probability. Along similar lines, Huang and Chen (2015) and Ren et al. (2018) assume customers resort to a heuristic, which involves sampling experiences of previous customers (referred to as anecdotal reasoning). In our model based on rational inattention, customers can also make mistakes. The main distinction is that customers decide on what they learn and to what detail, effectively controlling what type and extent of mistakes they make. This is analogous to customers forming optimal heuristics (Maćkowiak et al. 2018). Furthermore, customer behavior is adaptive to the business environment; changes in rewards, prices, delay cost, uncertainty as well as available time and attention all endogenously determine the amount of information processed and shape the resulting behavior. Our framework also fundamentally differs from bounded rationality models in its natural ability to connect the visible and invisible queues, through information costs.

The second stream of literature investigates the impact of additional queue size or delay information on queue joining behavior. One approach is to incorporate heterogeneity among customers in terms information they have about queue length. In a recent work, Hu et al. (2017) assume that an exogenously specified proportion of customers have perfect information about queue size, while the rest is completely uninformed about it (but they know the fraction of informed customers). They characterize the equilibrium and examine how heterogeneity impacts throughput and social welfare. Unlike Hu et al. (2017), customers are homogeneous in our model. But they optimally determine how much information they will acquire about the queue size. In this regard, we capture a desirable element that Hu et al. (2017) acknowledge as missing in their framework (see pp. 2654). We characterize and analyze the equilibrium that emerges from this micro-founded private learning efforts of customers.

An alternative approach assumes homogeneous customers can inspect the queue at a predetermined cost, upon which queue size is fully revealed. In a multi-server setting, Xu and Hajek (2013) assume customers inspect k out of N queues (randomly selected) at a fixed inspection cost linear in k and join the shortest one. The authors characterize the conditions of unique equilibrium. A closely related paper to ours that adopts the same approach is Hassin and Roet-Green (2017). They consider an invisible single-server service system where customers are allowed to inspect the queue and obtain full information at a fixed cost, leading to a model with three distinct decisions: join, balk, or inspect. The existence and uniqueness of an equilibrium strategy is proven and the effect of inspection cost on throughput and social revenue are discussed. We go a step further and allow the customers choose their information strategy and improve their knowledge on the state of the queue, which in optimality is never fully-informative (some uncertainty always remains). Customers can also decide to not acquire any information, if processing it is deemed to be not useful or too costly. Interestingly, this approach also avoids the complications caused in equilibrium analysis when customers make a choice between three distinct actions as in Hassin and Roet-Green (2017); once customers acquire information optimally, their eventual decisions is between joining or balking only.

We remark that both Hassin and Roet-Green (2017) and Hu et al. (2017) present results on how different levels of customer information might impact throughput and social welfare. In particular, the former shows that an intermediate level of inspection (information) cost can benefit throughput, while the latter shows that throughput can be unimodal in the proportion of informed customers.

We make the natural connections and draw parallels whenever possible. For example, consistent with earlier findings, we find that for high demand service systems, customers' limited information about the queue size can be beneficial for the firm (improve throughput). Our framework helps identifying the business conditions when this is most likely to happen. More importantly, unlike earlier studies, we are able to also identify cases where servicing customers with limited information has adverse effects on throughput; it first decreases and then increases in information costs (and levels). We achieve these insights through a unifying lens that combines customer inattention, endogenous information acquisition and strategic queueing behaviour in a natural and consistent manner.

It is worth noting that there are other papers that examine the effect of some form of delay information or information disclosure strategies from the service firm's perspective (e.g., Chen and Frank 2004, Dobson and Pinker 2006, Guo and Zipkin 2007, Economou and Kanta 2008, Simhon et al. 2016, among others). For a comprehensive survey of the literature related to delay announcements, we refer to Ibrahim (2018). We differ fundamentally from these works since there is no predetermined information disclosure strategy of the firm, rather customers obtain and process information that is optimal for them. Likewise, there are many other papers that study strategic queuing decisions, but with a different focus such as, queue size as a signal of quality (Debo et al. 2012, Kremer and Debo 2015), uncertain quality and queue choice (Veeraraghavan and Debo 2009), uncertain service rate and customer beliefs (Cui and Veeraraghavan 2016).

Our paper also contributes to the literature on rational inattention. With recent advances made in both theoretical and empirical grounds, there is a surge in the interest on rational inattention and its applications. Examples include, consumer (discrete) choice (Matějka and McKay 2015, Hüttner et al. 2018), pricing (Boyacı and Akçay 2017, Matějka 2015a, Matějka 2015b), energy efficiency (Sallee 2014), portfolio selection (Huang and Liu 2007), organizational focus (Dessein et al. 2016). To our knowledge, our paper is the first study that incorporates rational inattention in a service queueing setting with strategic customers.

3. Model

Consider a service system modeled as a basic single-server queue operating under FCFS (first-come-first-served) discipline, with Poisson arrival rate λ and exponentially distributed service times with mean $1/\mu$. Let R denote the unit reward a customer obtains upon being served and p denote the price charged by the firm. A customer arriving to the queue incurs a delay cost of C per time unit. Without loss of generality suppose that customers incur this cost only when they are waiting for the service (and not during service). Suppose that a customer arrives when there are n customers in the system. Then the pay-off, expected reward net of the delay cost, is given as

$$v_n \doteq R - p - \frac{C}{\mu} n. \tag{1}$$

Let the value of the outside option of balking (not joining) from the queue is normalized to 0. Clearly, if n is known, the customer will only join if $v_n \geq 0$. Alternatively, if all customers can observe the queue length freely, then they will only join if

$$n \leq n_e = \left\lfloor \frac{(R-p)}{C} \mu \right\rfloor. \tag{2}$$

This is essentially the threshold in Naor (1969) for visible queues. Let us assume that $R > p$ so that $v_0 > 0$, ruling out the uninteresting case where customers have no incentive to join the queue. In our setting, customers are not able to use this threshold policy because they are not able to discern the queue length precisely due to limited attention and cognitive capacity. We first describe how such customers would optimally acquire information and decide to join the queue or not. Subsequently, we characterize the equilibrium.

3.1. Join or Balk Decisions Under Limited Attention

Customers know that the number of customers ahead in the queue is a random variable and have a prior belief about its distribution (common to all). Let us denote the cdf of customers' prior belief as G and its pdf as g. Suppose for now that the belief distribution G is specified. One can view this as the anticipated queue size distribution, which in equilibrium will coincide with the actual distribution.

Rationally inattentive customers can ask questions and receive signals $\mathbf{s}$ to update their beliefs. Let ω denote the unknown state of the system at any time. The customer is free to select an *information processing strategy*, which is represented as the joint distribution $F(\mathbf{s}, \omega)$ of signals and states. The only requirement is that the marginal distribution over the states equals the prior distribution G, so that the customer's posterior beliefs are consistent with their priors. The customer chooses this distribution to maximize her *ex-ante* expected payoff minus the total cost of information $c(F)$ associated with generating signals of different precision levels. Following the works of Sims (2003) and Sims (2006), models of rational inattention quantify information acquisition and processing in terms of reduction in uncertainty, measured by the Shannon entropy. More specifically, let $H(B)$ denote the uncertainty of belief B measured by entropy. For a discrete distribution, $H(B) = -\sum_{\omega} P_{\omega} \log(P_{\omega})$ where P_{ω} is the probability of state the world $\omega \in \Omega$. Then the total cost of information associated with the information strategy F is given as

$$c(F) = \theta(H(G) - \mathbb{E}_{\mathbf{s}}[H(F(\cdot|\mathbf{s})]) \tag{3}$$

Here, $\theta > 0$ is the marginal cost of acquiring and processing information that the customer deems useful (simply referred to as cost of information hereon), and $F(\cdot|\mathbf{s})$ is the posterior belief about state after receiving the signal $\mathbf{s}$. Note that the total cost of information is defined as the *mutual information* between customer's prior and posterior beliefs, multiplied by the marginal information cost parameter θ. This cost function is well supported by information theory, since from Shannon's coding theorem it relates to the expected number of questions needed to be asked for implementing a particular information strategy (see Matějka and McKay 2015, Cover and Thomas 2012).

In the context of our queuing system, a customer has two discrete choices, either to "*join*" or to "*balk*". Given the prior belief G, she solves a two-stage problem. In the first stage, she selects an information strategy to refine her beliefs and in the second stage she selects the best option given her posterior belief. Let $V(B)$ denote the expected payoff from choosing the best option given some belief B and $\Omega = \{\omega_0, \omega_1, ...\}$ denote the state space of the total number of customers in the system. Then, a rationally inattentive customer's decision-making problem can be formally stated as:

$$\max_{F} \sum_{n=0}^{\infty} \int_{s} V\left(F\left(.|s\right)\right) F\left(ds|\omega_n\right) g\left(\omega_n\right) - C\left(F\right) \tag{4}$$

$$s.t. \int_{s} F\left(ds|\omega_n\right) = G\left(\omega_n\right) \text{ for } n \geq 0.$$

The first term in (4) is the ex-ante expected payoff from selecting the best option based on the generated posterior belief and the second term is the total cost of information given by (3). According to this model, the customer is optimally choosing (i) what and how much information to process (what to pay attention, how much attention to pay) and (ii) what action to select given the information.

A central result in rational inattention theory that helps to simplify the customer's problem is that each action can be selected in at most one posterior belief. In the context of queueing, this means that receiving distinct signals that lead to the same posterior is suboptimal, since it implies the acquisition of ample information that is not acted upon. The immediate consequence is that choosing signals is equivalent to choosing actions. As such, the mutual information between the signal and the state can be replaced with the mutual information between the chosen actions and state. Thanks to this property, it becomes possible to write an alternative maximization problem for the customer that uses state-dependent choices as decision variables, without any referencing to signals. Specifically, let S^J denote the set of signals that lead to joining decision. Then the induced *conditional* joining probability when there are n customers in the system can be represented $\pi_n = \int_{s\in S^J} F\left(ds|\omega_n\right)$. Let $\Pi = \{\pi_n; n \geq 0\}$ denote the collection of conditional joining probabilities; i.e. customer's joining policy. Based on this, it is also possible to write the *unconditional* joining probability as

$$\overline{\pi} = \sum_n \pi_n g_n \tag{5}$$

where g_n is the customer's prior probability about state ω_n. Then, for our queuing system, customer's equivalent optimization problem can be reformulated as

$$\max_{\Pi=\{\pi_n;n\geq 0\}} \sum_{n=0}^{\infty} v_n \pi_n g_n - C\left(\Pi, G\right) \tag{6}$$

$$s.t. \quad \pi_n \geq 0 \; \forall n \geq 0 \text{ and } \sum_{n=0}^{\infty} \pi_n = 1.$$

Here $\pi_n g_n$ is simply the joint probability that the customer joins and the state is ω_n. Since the utility of balking is normalized to 0, the first term is the total expected utility obtained under joining policy $\Pi = \{\pi_n; n \geq 0\}$. The second terms is the total cost of information $C\left(\Pi, G\right) = \theta\left(H\left(G\right) - E[H\left(G \mid \Pi\right)\right)$ quantifying the reduction in entropy, i.e. mutual information between the action and state, scaled by information cost θ. Due to symmetry of mutual information, $C\left(\Pi, G\right)$ can also be written as

$$\begin{aligned} C\left(\Pi, G\right) &= \theta\left(H\left(\Pi\right) - E\left[H\left(\Pi \mid G\right)\right]\right) \\ &= \theta\left(-\overline{\pi}\log\overline{\pi} - \left(1-\overline{\pi}\right)\log\left(1-\overline{\pi}\right) + \sum_{n=0}^{\infty} g_n\left(\pi_n\log\pi_n + \left(1-\pi_n\right)\log\left(1-\pi_n\right)\right)\right). \end{aligned} \tag{7}$$

It is established in the rational inattention literature that the optimal information processing strategy for any $\theta > 0$ results in conditional choice that follows a generalized multinomial logit (GMNL) formula (Matějka and McKay 2015). In particular, when the choice is about joining a queue or not as described above, the conditional probability π_n of joining the queue when there are n customers satisfy

$$\pi_n = \frac{\overline{\pi} e^{v_n/\theta}}{\overline{\pi} e^{v_n/\theta} + 1 - \overline{\pi}} \quad \text{almost surely for } \theta > 0 \tag{8}$$

where $\overline{\pi}$ is the *unconditional* probability of joining the queue that needs to satisfy equation (5) for consistency. If $\theta = 0$, then with probability 1, the customer joins or balks deterministically, depending on which one yields the higher payoff in that state.

The conditional joining probabilities characterized by the GMNL equation (8) capture the intricate relationship between three central drivers of customer's decision, namely the payoffs, beliefs, and information costs. Observe first that according to (8) if the customer has a positive probability of joining in at least one state (i.e., $\overline{\pi} > 0$), then she has a positive probability to join in all other states of the system. However, the higher is her payoff (v_n), the more likely she will join. In our queuing system, this will also imply that state-dependent joining probabilities increase as n decrease (as v_n increase). The impact of prior beliefs are captured through the unconditional probability $\overline{\pi}$. It is crucial to note that $\overline{\pi}$ is *not* an exogenous parameter; rather it is part of customer's endogenous decision making process. One needs to solve the above optimization problem together with (8) to arrive at a complete explicit solution. Rewriting (8) as $\pi_n = \left(e^{v_n/\theta + ln(\overline{\pi})}\right) / \left(e^{v_n/\theta + ln(\overline{\pi})} + 1 - \overline{\pi}\right)$ it is evident that unconditional probability $\overline{\pi}$ effectively shifts the customer's payoff. Hence, her joining decision is swayed by how "attractive" it is a-priori to join the queue. Information costs play a strong role on how much emphasis the customer puts on the beliefs. When θ is low, the customer can acquire more information about each state and the payoff. In the limit, the queue is visible, and she deterministically makes the best choice in each state. In contrast, as θ increases, she acquires less information and relies more on her belief. In the limit, the queue is invisible, and the customer deterministically joins or balks based on her ex-ante beliefs.

Continuing with the analysis, we can plug the conditional joining probability π_n given by (8) into customer's optimization problem in (6), which yields a simplified representation:

$$\max_{\overline{\pi} \in [0,1]} \theta \mathbb{E}_G \left[log \left(\overline{\pi} e^{v_n/\theta} + 1 - \overline{\pi} \right) \right]. \tag{9}$$

Hence, in effect, the customer is choosing the unconditional probability $\overline{\pi}$. It is clear that problem in (9) is concave in $\overline{\pi}$ with linear constraints and can be easily solved. Once unconditional joining probability $\overline{\pi}$ is found, conditional joining probabilities are calculated using (8).

It is worth noting that equation (8) constitute the necessary conditions for optimality but they are not sufficient. For instance the queueing policy $\Pi_1 = \{\pi_n = 1; n \geq 0\}$ where everyone joins and the policy $\Pi_0 = \{\pi_n = 0; n \geq 0\}$ where everyone balks at each state automatically satisfy these conditions. Hence, it holds trivially when actions (join or balk) are not chosen at all, but does not specify when this may occur. The next lemma presents the complete characterization of customer's optimal joining policy, including both the necessary and sufficient conditions. The proofs for the next and subsequent results are relegated to the appendix.

Lemma 1 **(Necessary and Sufficient Conditions)**. *The policy* $\Pi = \{\pi_n; n \geq 0\}$ *is optimal if and only if the implied unconditional choice probabilities* $\overline{\pi} = \sum_n \pi_n g_n$ *for actions Join and Balk satisfy*

$$\sum_{n=0}^{\infty} \frac{e^{v_n/\theta} g_n}{\overline{\pi} e^{v_n/\theta} + 1 - \overline{\pi}} \leq 1 \qquad \textit{(for “Joining”)} \tag{10}$$

$$\sum_{n=0}^{\infty} \frac{g_n}{\overline{\pi} e^{v_n/\theta} + 1 - \overline{\pi}} \leq 1 \qquad \textit{(for “Balking”)} \tag{11}$$

and both equations (10) *and* (11) *have to hold with equality if* $0 < \overline{\pi} < 1$*. Otherwise, the sufficient conditions are*

$$\mathbb{E}_G \left[e^{v_n/\theta} \right] \leq 1 \textit{ for } \overline{\pi} = 0, \tag{12}$$

$$\mathbb{E}_G \left[e^{-v_n/\theta} \right] \leq 1 \textit{ for } \overline{\pi} = 1. \tag{13}$$

Lemma 1 establishes that there are cases that yield join or balk decisions with certainty, i.e., without the need for processing any information. For instance it is possible that condition in (12) is satisfied when the customer's prior belief on low states (i.e., when there are few customers in the queue) is very weak and it is optimal for the customer to balk with certainty, i.e., $\overline{\pi} = 0$. A similar effect may also take place when the customer attaches a very low value to the service provided, i.e., a low service reward. On the contrary, when the customer strongly believes that there will be few customers in the queue and/or her reward from service is high enough, then she may decide to join with certainty without obtaining further information, i.e., $\overline{\pi} = 1$ and (13) is satisfied.

At this point it is also worthwhile to make connections with more traditional search models. A prominent example is the model by Hassin and Roet-Green (2017), where arriving customers have the opportunity to buy perfect information at a fixed cost. In our framework, receiving perfect information is equivalent to reducing the posterior entropy $\mathbb{E}_{\mathbf{s}}[H(F(\cdot|\mathbf{s})]$ to zero. This is tantamount to customers paying a fixed cost of $\theta H(G)$ and deciding to join based on the current queue size (or not pay and decide on the basis of prior belief G). However, this is a suboptimal strategy when information strategy is endogenized. Generating perfect signals is never optimal since this is prohibitively expensive (leads to an unbounded information cost, as per (3)). This is why the optimal decision is always probabilistic if the customer chooses to acquire and process information. Nevertheless, the customer can also choose not to process any information (see Lemma 1).

3.2. Queueing Behaviour in Equilibrium

Until this point, we have assumed that customers have exogenously specified prior beliefs, representing the anticipated queue distribution. In fact customers are strategic in our framework; they are aware of the other rationally inattentive customers and anticipate their actions. As a result, customer beliefs are formed by queuing behavior in equilibrium. Queuing behavior itself is shaped by the optimal information acquisition and joining decisions of rationally inattentive customers in each state. Next, we define such an equilibrium. To this end, let $\tilde{G}$ and $\tilde{g}$ denote the cumulative and probability distribution functions for the number of customers in the system in equilibrium and $\rho = \lambda/\mu$ is the utilization factor.

Definition 1. In the queueing system with rationally inattentive customers with information cost $\theta > 0$, the equilibrium probability of joining the queue when there are n customers present is

$$\widetilde{\pi}_n = \frac{\widetilde{\pi} e^{v_n/\theta}}{1 - \widetilde{\pi} + \widetilde{\pi} e^{v_n/\theta}}, \quad \text{where} \tag{14}$$

$$\widetilde{\pi} = \sum_n \widetilde{\pi}_n \tilde{g}(n) \tag{15}$$

is the equilibrium unconditional joining probability, i.e., equilibrium fraction of customers joining the queue. Denoting $\lambda_n = \lambda\widetilde{\pi}_n$ as the state-dependent equilibrium arrival rates, the equilibrium steady-state distribution $\tilde{g}$ is characterized as

$$\tilde{g}_0 = \frac{1}{1+\sum_{k=1}^{\infty}\frac{\lambda_0\lambda_1\ldots\lambda_{k-1}}{\mu^k}} \quad \text{and} \quad \tilde{g}_n = \tilde{g}_0\frac{\lambda_0\lambda_1\ldots\lambda_{n-1}}{\mu^n} \quad \text{for } n \geq 1.$$

The more critical question is whether such an equilibrium exists. In the next theorem, we show that it indeed does for any $\theta > 0$ and it is unique.

Theorem 1. *There exists a unique equilibrium satisfying Definition 1.*

Although our framework involves nontrivial customer behaviour in terms of queueing, Theorem 1 establishes a strong result that a unique equilibrium exists despite the complexity of the model. Finding this equilibrium requires solving a fixed point equation since it requires a consistency between the joining probabilities (which is a solution of the rational inattentive customer's optimization problem) and resulting queue distribution (which is an input to the same optimization problem).

There still remains the question of whether rationally inattentive customers can form the correct beliefs about the queue size distribution, i.e., whether the equilibrium can be attained. We show in the appendix that such an equilibrium can be attained via adaptive learning in a setting where customers can observe the joining behaviour of past customers. More specifically, consider a model with multiple periods and suppose that in each period, customers form their beliefs about the queue size distribution based on the average of joining probability of past periods, and then make the optimal joining decision contingent on that belief in a rationally inattentive manner. In a finite number of periods, joining probabilities converge to the equilibrium in Definition 1.

An immediate corollary of Theorem 1 is the limiting cases of the information cost, which bridge the two extreme scenarios in strategic queueing, namely visible and invisible queues.

Corollary 1. *The following are true:*

1. *(Visible queues) When $\theta = 0$, customers can determine the queue length exactly and decide to join only if the number of people in the system is less than or equal to the critical number n_e defined in* (2)*. Corresponding equilibrium joining fraction is*

$$\widetilde{\pi} = 1 - \frac{\rho^{n_e+1}}{1 + \sum_{k=1}^{n_e+1} \rho^k} = \frac{1 - \rho^{n_e+1}}{1 - \rho^{n_e+2}}. \tag{16}$$

2. *(Invisible queues) When $\theta = \infty$, customers base their queuing decisions on their prior beliefs only, resulting in the following outcomes (when $\lambda < \mu$)*
   (a) *(Always join) If $R - p - \frac{C\lambda}{\mu(\mu-\lambda)} \geq 0$, then all customers join the queue.*
   (b) *(Mixed strategy) Otherwise, unique equilibrium joining fraction is $\widetilde{\pi} = \left[\rho\left(1 + \frac{C}{\mu(R-p)}\right)\right]^{-1}$.*

Complete characterization of equilibrium joining fraction when $\theta = \infty$ is;

$$\widetilde{\pi} = \min\left\{\widetilde{\pi} = \left[\rho\left(1 + \frac{C}{\mu(R-p)}\right)\right]^{-1}, 1\right\}. \tag{17}$$

The two extreme cases of our framework covered in Corollary 1 retrieve the classical models and results in the literature. The first case is precisely the scenario with visible queue model of Naor (1969), where the equilibrium is a threshold policy. The latter case is precisely the invisible queue model of Edelson and Hilderbrand (1975). In particular, if the benefit of joining the queue is positive even if everyone else joins, then customers join with probability 1. Otherwise, customers join with a fixed probability. Note that given our assumption that $R > p$, the case where all customers balk does not occur.

Next, we investigate the impact of salient characteristics of service systems, including the reward from service R, price p, delay cost per unit time C, and service rate μ on customer joining behaviour $\widetilde{\pi}$ (or equivalently system throughput $\lambda\widetilde{\pi}$).

Proposition 1. *Equilibrium joining fraction $\widetilde{\pi}$ and throughput increases in R and μ and decreases in p and C.*

Intuitively, customer joining probability in equilibrium should increase in the utility customers obtain from joining. This is surely the case when the reward from service R is higher, and the price p and/or waiting cost per unit time C are lower. For this reason, we define $\overline{R}=(R-p)/C$ as the service attractiveness. Proposition 1 confirms that equilibrium joining probability improves with service attractiveness. The effect of service rate μ, on the other hand, is not as straightforward. Although a faster service may potentially mean less congestion, it also incentivizes customers to join which in turn may increase congestion. It turns out that the first affect is always stronger, and a faster service rates yields more joining customers in equilibrium.

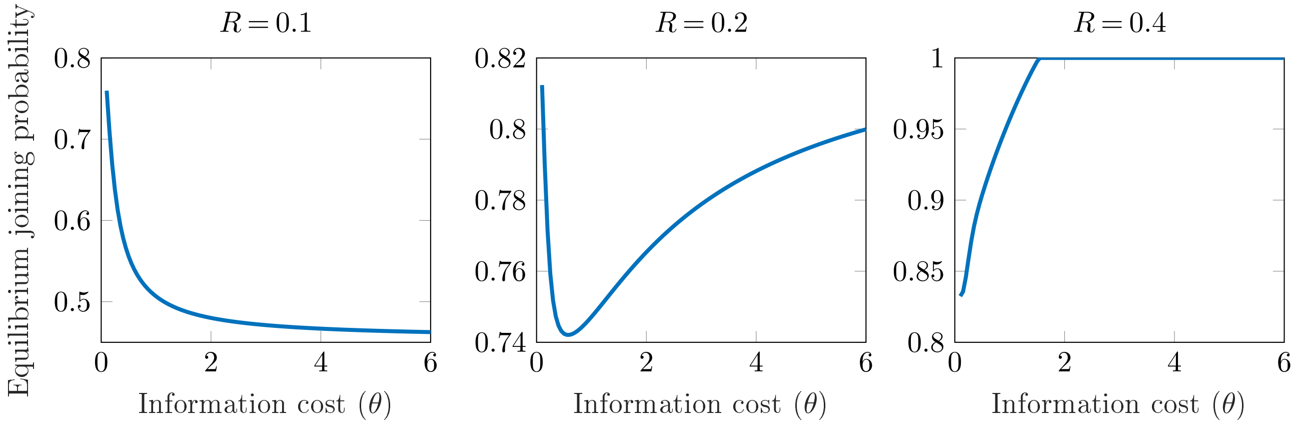
The impact of the information cost on queuing behavior is more subtle and cannot be analytically ascertained. We take a deeper look into it through numerical experiments in the next section.

4. Impact of Information Cost on Throughput

Information costs impact the extent of learning customers can afford to (or able to) undertake regarding queue size. At a first glance, it seems quite plausible that the effects of θ should be monotonic, and its direction should depend whether in equilibrium visible or invisible queues have higher joining fractions. This intuition turns out to be only partially correct. To illustrate this, in the following discussion, we use $\widetilde{\pi}^{(\theta)}$ to denote the equilibrium joining fraction when the unit information cost is θ.

We find that the impact of information cost on throughput is governed by both the level of demand (congestion) for service and the attractiveness $\overline{R}$ of the service to the customers. Specifically, we observe that for any demand level, there is a range for service attractiveness in which throughput is either decreasing-increasing or increasing-decreasing in information cost θ. If $\overline{R}$ is below this range, then throughput is decreasing in θ, while if $\overline{R}$ is beyond this range, then throughput is increasing in θ. Furthermore, the range where this non-monotone impact is observed also depends on the level of demand for the service.

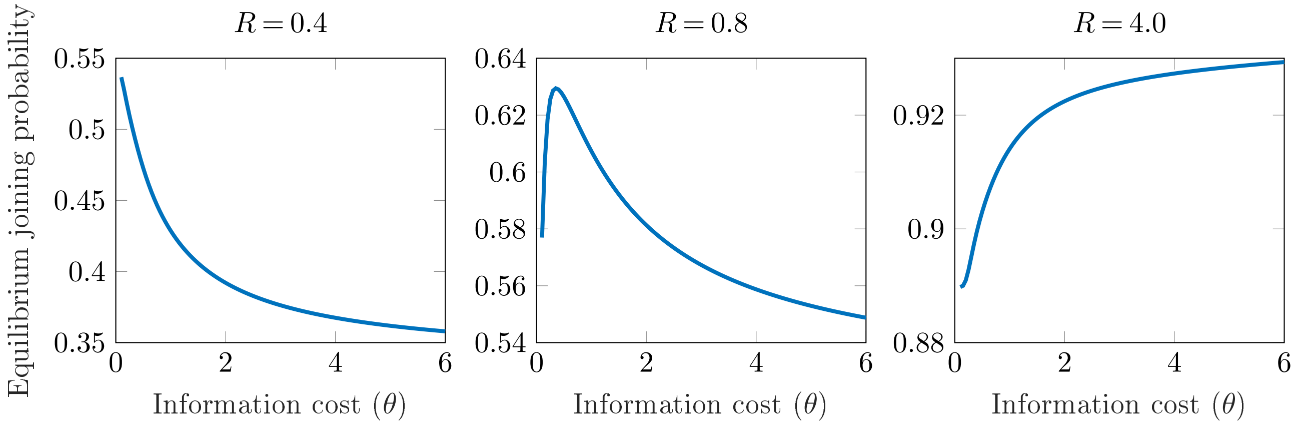
In order to see this, note that when service is quite attractive to the customers, they will tend to join without processing information (based on beliefs), and hence equilibrium joining fraction for an invisible system will be very high. When the queue is visible, however, some customers will still not join due to congestion. In such cases, $\widetilde{\pi}^{(0)}$ is relatively low compared to $\widetilde{\pi}^{(\infty)}$, and equilibrium joining fraction $\widetilde{\pi}^{(\theta)}$ is monotonically increasing in θ. In contrast, when service is quite unattractive, customers are unlikely to join an invisible queue, $\widetilde{\pi}^{(\infty)}$ will be low. For a visible queue however, some (lucky) customers will still be able to find the queue short enough to warrant joining. In such cases, $\widetilde{\pi}^{(0)}$ is relatively higher compared to $\widetilde{\pi}^{(\infty)}$, and equilibrium joining probability $\widetilde{\pi}^{(\theta)}$ is monotonically decreasing in θ. When the equilibrium joining fractions under visible and invisible queues are not very distinct, then we observe quite different, and possibly non-monotone behavior with respect to information costs. This occurs within an intermediate range of service attractiveness $\overline{R}$. We elaborate more on these ranges in the following two examples for low demand and high demand scenarios respectively.



(a) Low service attractiveness (b) Mod. service attractiveness (c) High service attractiveness

Figure 1 Impact of information cost on equilibrium joining fraction and throughput: low demand ($\lambda = 0.2, \mu = 1, C = 1$)

Figure 1 illustrates three different cases when the demand rate is low. Evidently, when service attractiveness $\overline{R}$ is very low, higher information costs lead to lower throughput. Interestingly, when $\overline{R}$ is intermediate (but still relatively low), throughput first decrease and then increases. The intuition is as follows. When $\theta = 0$, customers can observe the queue length and deterministically join. When information cost θ is slightly increased, customers process information and due to their limited attention, they may not be able to discern queue size when it is in fact relatively short, and balk. This has a negative impact on throughput. Arguably, for the same reason, customers may not discern queue size when it is longer, and erroneously join instead of balking. However, this is less likely to happen because of low demand and congestion. Hence, initially throughput decreases in θ. When information cost is further increased and customers start to weigh more their beliefs, they join with a higher probability because they believe the system is not congested and despite the low reward it makes sense to join instead of balking. As a result, throughput starts increasing in θ. Finally, when service attractiveness is high enough, as explained earlier, throughput is monotonically increasing in θ.



(a) Low service attractiveness (b) Mod. service attractiveness (c) High service attractiveness

Figure 2 **Impact of information cost on equilibrium joining fraction and throughput: high demand** ($\lambda = 0.85, \mu = 1, C = 1$)

Figure 2 illustrates three different cases when the demand rate is high. As in the case of low demand, there is an intermediate level of service attractiveness that makes throughput non-monotone in information costs. Strikingly, here the effect is opposite; throughput first increases and then decreases, i.e., unimodal. The intuition follows a logic very similar to the low demand case, but the effects are reversed. This is because when demand is high, longer queue sizes are more likely, and therefore (erroneous) joining decisions at higher queue lengths due to limited attention outnumber (erroneous) balking decisions at lower queue lengths. Furthermore, for higher demand levels, the intermediate range for which non-monotone behavior is observed is wider, and occurs for higher information cost levels. Nevertheless, as before, when service attractiveness is outside of this range, information costs have a uniform effect and either decrease or increase throughput (Figures 2a,2c). The above pattern persists for moderate demand levels as well. It is also possible to observe a combination pattern with decreasing-increasing followed by a decreasing throughput in this range.

Our numerical investigation demonstrates clearly that the complex, non-monotone behavior prevails when demand and service attractiveness are strongly coupled (one can also think of it as *positive correlation*). We can substantiate this by focusing on a case where the equilibrium joining fractions for visible and invisible queues are identical. To this end, suppose that the threshold $n_e = 0$. Then, $\widetilde{\pi}^{(0)} = \widetilde{\pi}^{(\infty)}$ whenever $(1+\rho)^{-1} = [\rho\,(1 + C/\,(\mu(R-p))]^{-1}$, or equivalently; $\rho = (R-p)\mu/C$. Normalizing $\mu = 1$, it becomes evident that equilibrium joining fractions under visible and invisible queues are the same when $\lambda = \overline{R} = (R-p)/C$, i.e., when demand and service attractiveness display perfect positive correlation. Figure 3 illustrates the effect of θ on $\widetilde{\pi}^{(\theta)}$, for various values of λ for this case. Note that as demand increases, equilibrium joining probability starts to turn from “first decreasing, then increasing” to “first increasing, then decreasing”.

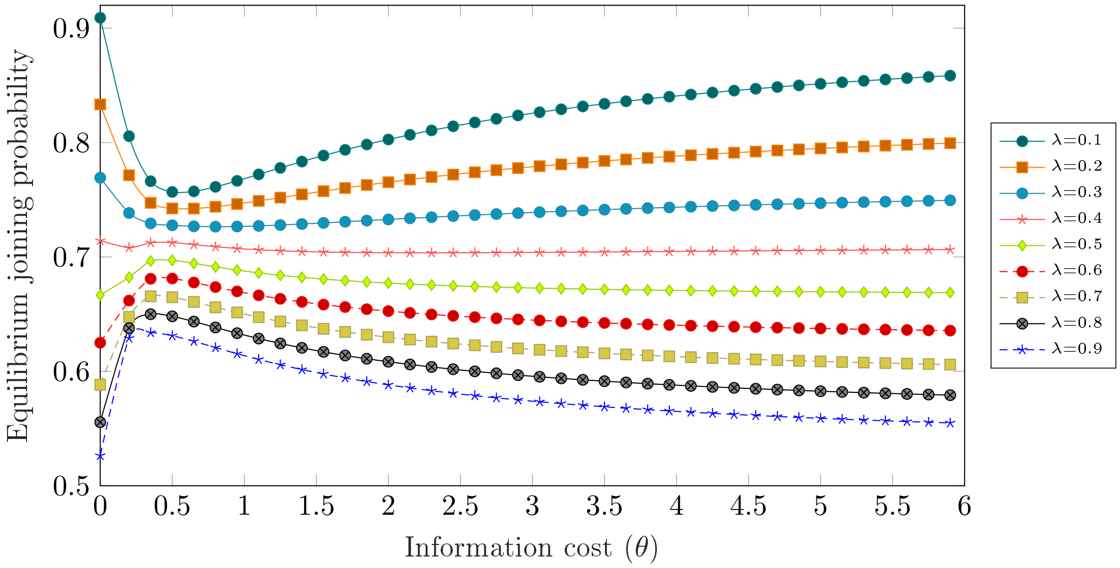


Figure 3 **Impact of information cost on equilibrium joining fraction when** $\widetilde{\pi}^{(0)} = \widetilde{\pi}^{(\infty)}(C = 1, \mu = 1, R = \lambda)$

Here we also remark that there are some other papers that find that throughput might be unimodal in terms of information prevalence. Hassin and Roet-Green (2017) conclude that a positive and finite inspection cost might achieve a higher throughput. Similarly, Hu et al. (2017) finds that having a portion of uninformed customers in the society might be better in terms of throughput. Our framework is able to explain these results using a simple model that is rooted in the first principles and systematically links the main drivers of human decision making such as beliefs, payoffs and information costs. More importantly, our framework is also able to explain counter cases where throughput suffers from limited attention alluding to the potential dangers of deliberate obstruction of information acquisition. To the best of our knowledge, this is not noted in the strategic queueing literature.

Our results offer a compounding insight on the effect of information prevalence on customer behaviour and hence have significant managerial implications for service firms. It establishes that when customers highly value a service that is already popular, the firm can benefit from an opaque queue and find in its best interest to make information acquisition difficult or even to deliberately obstruct it to some extent, rather than providing a completely visible or invisible system. The opposite is true when the firm faces low demand from customers who do not value the service much. In this case, it is optimal for the firm to employ an "all-or-nothing" kind of an information strategy in terms of "queue transparency".

5. Social Welfare

We now investigate how customers' rational inattention affects social welfare. For a given price p and information cost θ, social welfare is defined as the expected net utility of society (including customers and service provider) per unit time. In other words, it is the sum of consumer and firm surplus. Note that information cost is a real cost that customers incur so it is included in social welfare. Service fee p, on the other hand, is merely a transfer payment between customer and service provider, and therefore does not directly impact social welfare. However, it still has a strong impact through the queueing joining behaviour it induces. Defining $L_q = \sum_{n=1}^{\infty}(n-1)\widetilde{g}_n$ as the expected number of customers in the queue at a given time, we can write the social revenue as:

$$W = \lambda\widetilde{\pi}R - \lambda C\left(\widetilde{\Pi}, \widetilde{G}\right) - CL_q \tag{18}$$

where total information cost for a customer $C\left(\widetilde{\Pi}, \widetilde{G}\right)$ is defined in (7).

When $p = 0$, social welfare only consists of customer surplus in equilibrium. In this case, we observe that social welfare is decreasing in information cost θ. This is fairly intuitive as customers benefit from making more informed decisions at a lower cost and both capacity and demand are better matched. This results in less congested queues where every customer benefits. This finding is also largely consistent with social welfare results in the literature (Hu et al. 2017, Hassin and Roet-Green 2017). In Hassin and Roet-Green (2017), service fee is zero, and social welfare decreases as the cost of inspecting the queue increases. In Hu et al. (2017) social welfare increases as the proportion of informed customers in the population increases. An exception occurs when uninformed customers choose to balk with certainty,

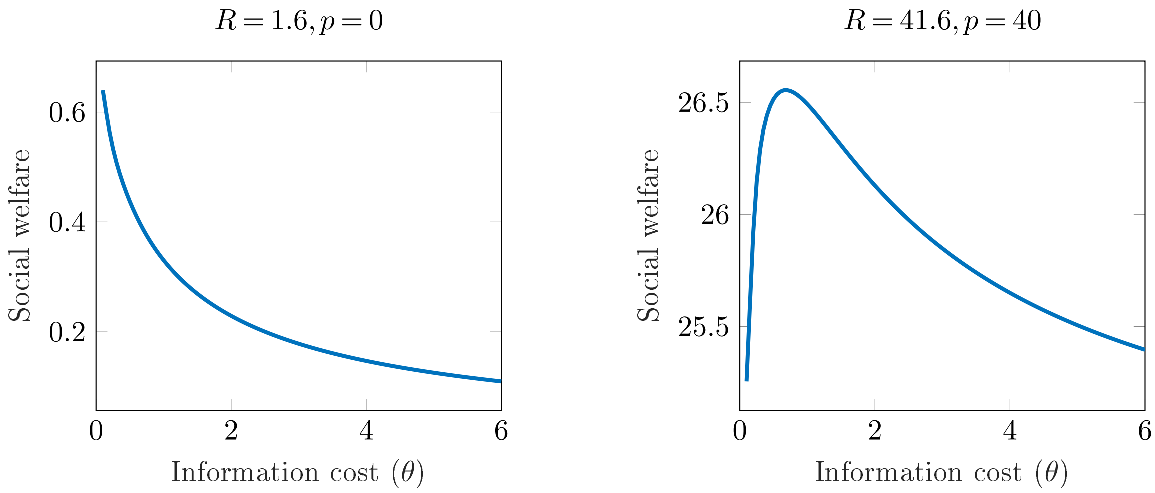


Figure 4 **Impact of information cost on social welfare** ($c = 1, \mu = 1, \lambda = 0.85$)

but in this case the market consists only of informed customers and increasing their proportion is mainly about increasing demand and not about information prevalence.

When $p \neq 0$ and firm surplus is taken into account, social welfare may not exhibit monotonicity in terms of information prevalence. Indeed, we know that the firm may benefit from limited attention and information costs. In such cases, the decrease in customer surplus due to increased information cost may not offset the increase in firm surplus. To see this clearly, let us disentangle social welfare into its two components, consumer and firm surplus. Let $W(R,p,\theta)$ and $\widetilde{\pi}(R,p,\theta)$ denote the social welfare and equilibrium joining fraction for consumers when service reward is R, fee is p and information cost is θ. Then we can write social welfare as the sum of social welfare that would have been obtained with effective service reward $R-p$ and and zero service fee, and firm surplus, i.e., $W(R,p,\theta) = W(R-p,0,\theta) + \lambda\widetilde{\pi}(R-p,0,\theta)p$. Note that $\widetilde{\pi}(R,p,\theta) = \widetilde{\pi}(R-p,0,\theta)$. Mathematically speaking, social welfare is the sum of a decreasing function (consumer surplus) and a potentially non-monotone function (firm surplus). Hence, it is possible to generate different social reward behaviour by keeping $R-p$ constant and increasing p. An illustration is provided in Figure 4 where the non-monotone equilibrium joining probability is coupled with a high service fee p to generate a unimodal social welfare function. Hence, if a popular service that is highly valued by customers is also very profitable for the firm, the total social welfare might benefit from information frictions.

6. Revenue Maximization

We now examine revenue maximization from the firm's perspective. We assume that the firm charges a single, state-invariant price to maximize its revenue per unit time and solves

$$\max_{p\geq 0} R_I(p) = \lambda p \widetilde{\pi}(p) \tag{19}$$

where $\widetilde{\pi}(p)$ is the equilibrium unconditional joining probability as a function of price. In the following proposition, we show that revenue function in (19) is unimodal in price.

Proposition 2. *$R_I(p)$ is unimodal in p and there exists a unique maximizer $p^* \geq 0$.*

We explore the behaviour of revenue-maximizing price and corresponding optimal firm revenue with respect to information cost numerically. First, we observe that the optimal price and revenue can be quite erratic. Second, the nuanced impact of information cost on customer joining behaviour (hence throughput) remain prominent when the firm optimally sets the price. More specifically, when both service attractiveness and potential demand are high, customers join more often when queue is opaque. (see Figures 2b,3). Given this tendency, the firm can afford to increase price and extract a premium from the customers. In contrast, when both demand is low and service is not attractive, it is optimal for the firm to reduce to price and moderate the customer losses due to limited attention. Nevertheless, the optimal price and revenue display the same non-monotone behavior of the throughput in both cases (albeit in narrower ranges due to pricing). An illustration is provided in Figure 5.

A final remark is in order here regarding the comparison of firm-optimal and socially optimal (i.e., social welfare maximizer) prices. It is known that for visible queues, a profit-maximizing firm charges a higher price than socially optimal. On the other hand, for invisible queues firm's optimal price is also socially optimal (Naor (1969), Edelson and Hilderbrand (1975)). Our model with inattentive customers also retrieves these results. We also observe that for any finite information cost $\theta \geq 0$ the firm continues to charge a higher price than socially optimal, and as θ approaches infinity, the two prices converge to the same value, conforming the equivalence result for invisible queues. We omit details for brevity.

7. Versatility of the Framework

Our framework can be easily modified to include different service system characteristics and customer behavior. We now present some variations and extensions, and highlight their potential applications.

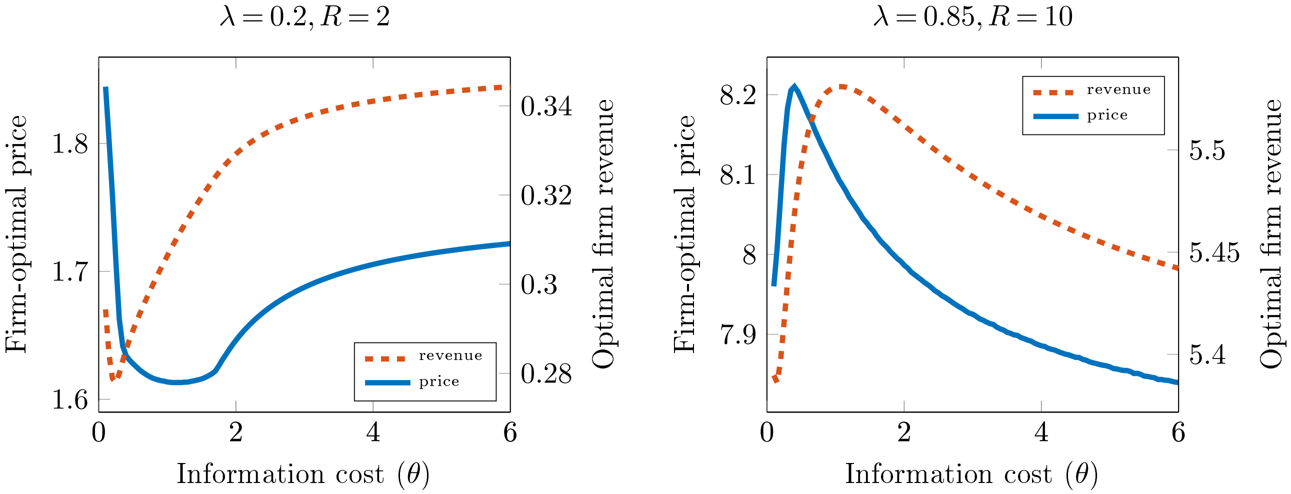


Figure 5 **Impact of information cost on firm-optimal price and revenue ($\mu = 1, C = 1$)**

7.1. Finite Queue Capacity

Consider a service system with a finite waiting capacity N. An arriving customer is not admitted to the system when it is full, and to reflect this inconvenience, there's a rejection cost $T \geq 0$. The state space (total number of customers) for this queueing system is then $\{0, 1, .., N\}$. Similar to the baseline model, the value of not joining is normalized to 0, while the pay-off in each stage $n = 1, .., N-1$ is given by (1). For state N, $v_N = -T$. The equilibrium of the finite capacity queueing system can be defined analogous to Definition 1, with the only change being the limited state space $\{0, 1, .., N\}$. Note that this framework can also be readily adapted for a service system with multiple servers. In both cases, it can be verified that a unique equilibrium strategy exists.

The baseline model we analyzed corresponds to the limiting case with $N = \infty$. In order to shed light on the impact of finite queue capacity, we elaborate on the other limiting scenario with zero waiting capacity, $N = 1$. This implies that arriving customers can not join when the server is busy. In this model, customers aim to learn whether the server is busy or not to make a "joining" decision. Since they are rejected when the queue is full, these are rather "trying" decisions. The states for this queueing system is clearly $\{0, 1\}$ and the equilibrium is characterized as follows.

THEOREM 2. *In a queueing system with no waiting capacity and rationally inattentive customers with information cost $\theta > 0$, the unique equilibrium unconditional joining (trying) probability $\widetilde{\pi}$ is*

$$\widetilde{\pi} = \min\left\{ \frac{\left(e^{v_0/\theta} - 1\right)}{\left(1 - e^{v_1/\theta}\right)\left(\rho e^{v_0/\theta} + e^{v_0/\theta} - 1\right)}, 1 \right\}. \tag{20}$$

The conditional joining probabilities in equilibrium are $\widetilde{\pi}_n = \left(\widetilde{\pi} e^{v_n/\theta}\right) / \left(\widetilde{\pi} e^{v_n/\theta} + 1 - \widetilde{\pi}\right)$, for $n \in \{0, 1\}$ and the equilibrium steady-state probability of the server being idle is $\widetilde{g}_0 = (1 + \rho\widetilde{\pi}_0)^{-1}$.

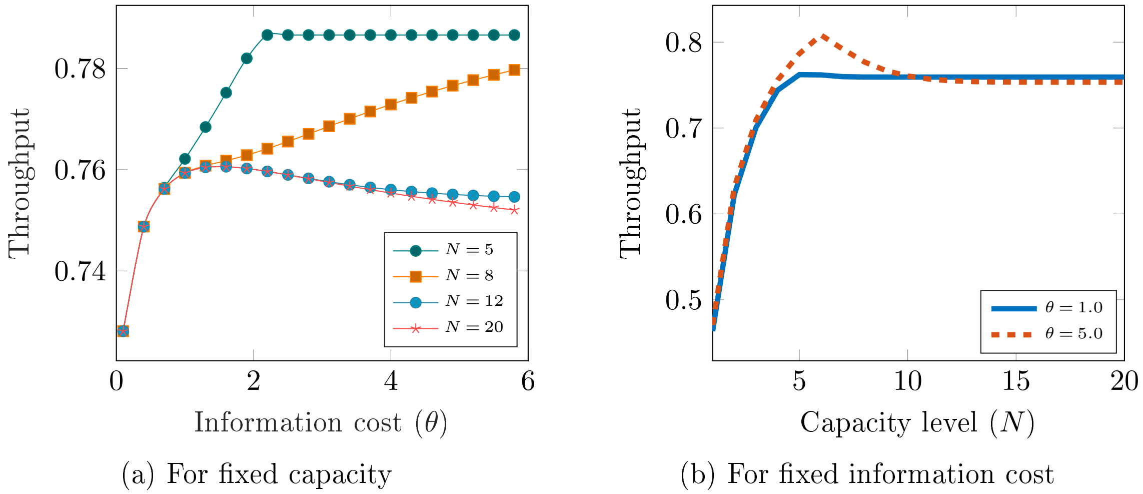
Throughput in equilibrium is defined as $\lambda \sum_{k=0}^{N-1} \widetilde{\pi}_n \widetilde{g}_n$ for a service system with capacity N. It is clear that throughput is not proportional to the equilibrium joining fraction since there is no entrance in state N. Next proposition characterizes the effect of information cost on both equilibrium joining probability and throughput for the zero-capacity case.

PROPOSITION 3. *The following are true:*

(1) If $R - p \geq T$, the equilibrium joining fraction $\widetilde{\pi}$ is increasing in information cost θ. Otherwise, $\widetilde{\pi}$ takes its maximum value when $\theta = 0$.

(2) Throughput takes its maximum value when $\theta = 0$.

Proposition 3 states that in a service system with no waiting room, the firm should make the system completely visible to maximize throughput, regardless of the rejection cost to the customers. The rationale is that obstructing information acquisition for customers can only deter them joining the system when it is empty, which is definitely undesirable for the firm. At the other extreme with infinite waiting room, we already know that throughput may exhibit non-monotone behavior. Putting these together, it is evident that queue capacity can be an important design consideration when customers have limited attention. This is corroborated in Figure 6a, which depicts the impact of information costs on throughput for different queue capacity levels ($N = 20$ is practically identical to our baseline case). Most notably, our extended framework elucidates the possibility for the firm to benefit from *limiting* waiting room capacity. As evident in Figure 6b, throughput may be maximized at an intermediate, finite waiting room capacity.



(a) For fixed capacity

(b) For fixed information cost

Figure 6 **Impact of finite capacity level on customer behaviour** ($R = 2.8, p = 0, C = 1, \mu = 1, \lambda = 0.9, T = 2$)

7.2. Environmental Uncertainty

There could be aspects of the service system other than the queue size that are not easily discernible by customers, such as service speed, quality and reward, among others. Consider an extension of our framework where the scope of "environmental uncertainty" is expanded to include both service value R (quality of service) and service rate μ. Suppose that customers optimally allocate their attention to learn about the uncertain service reward and service speed, along with the uncertain queue size. Let us assume that R and μ are discrete random variables with finite support; state space for R consists of K_R distinct values $\Omega_R = \{R_k : k = 1, .., K_R\}$ and state space for μ consists of K_μ distinct values $\Omega_\mu = \{\mu_k : k = 1, .., K_\mu\}$ both in ascending order. State of the queuing system at any time is the a triplet $\Omega = \{X = (N, R, \mu) : N \geq 0, R \in \Omega_R, \mu \in \Omega_\mu\}$ with prior joint distribution $h(.)$. We use $h_{\mathcal{X}}$ to denote the marginal probability of $\mathcal{X} \subset \Omega$. Both Ω and h are assumed to be common knowledge in the population.

Note that queue size distribution depends on service rate realization. Hence, we need to define state-dependent queue distributions in equilibrium. In particular, let $\tilde{g}(j) = \{\tilde{g}_n(j); n \geq 0, j = 1, .., K_\mu\}$ denotes the conditional steady-state queue distribution in equilibrium given service rate μ_j. We also use $\widetilde{\pi}_n(i, j)$ to denote the conditional probability of joining in equilibrium when service reward is R_i, service rate is μ_j and number of customers in the system is $n \geq 0$. State-dependent utility of joining is denoted as $v_n(i, j) = R_i - p - cn/\mu_j$ and value of balking is normalized to zero. For a given prior h, rationally inattentive customers solve the extended version of the optimization problem (9) to arrive at the optimal unconditional joining probability: $\max_{\overline{\pi}\in[0,1]} \left[\theta \sum_{n=0}^{\infty} \sum_{i=1}^{K_R} \sum_{k=1}^{K_\mu} h\left(n,i,j\right) log \left(\overline{\pi} e^{v_n(i,j)/\theta} + 1 - \overline{\pi}\right)\right]$. The difference is that the expectation is now taken with respect to the generalized joint distribution h. We define the equilibrium as follows:

Definition 2. In the queueing system with rationally inattentive customers with information cost $\theta > 0$, the equilibrium probability of joining the queue when there are n customers present, service reward is R_i and service rate μ_j is

$$\widetilde{\pi}_n\left(i,j\right) = \frac{\widetilde{\pi} e^{v_n(i,j)/\theta}}{1 - \widetilde{\pi} + \widetilde{\pi} e^{v_n(i,j)/\theta}} \text{ for } i \in \left\{1,..,K_R\right\}, j \in \left\{1,..,K_\mu\right\}, n \geq 0$$

where $\widetilde{\pi} = \sum_{n=0}^{\infty} \sum_{i=1}^{K_R} \sum_{k=1}^{K_\mu} \widetilde{\pi}_n\left(i,j\right) \tilde{g}_n\left(i,j\right) h_{R,\mu}\left(i,j\right)$ is the unconditional joining probability in equilibrium. Denoting $\lambda_n\left(i,j\right) = \lambda \widetilde{\pi}_n\left(i,j\right)$ as the state-dependent arrival rates, the equilibrium conditional steady-state distribution given service reward R_i and service rate μ_j is

$$\tilde{g}_0\left(i,j\right) = \frac{1}{1 + \sum_{n=1}^{\infty} \frac{\lambda_0(i,j)\lambda_1(i,j)...\lambda_{n-1}(i,j)}{\mu_j^n}},$$
$$\tilde{g}_n\left(i,j\right) = \tilde{g}_0\left(i,j\right) \frac{\lambda_0\left(i,j\right)\lambda_1\left(i,j\right)...\lambda_{n-1}\left(i,j\right)}{\mu_j^n} \quad \text{for } n \geq 1.$$

Resulting conditional joining probabilities $\widetilde{\pi}_n\left(i,j\right)$ form the state dependent arrival rates $\lambda_n\left(i,j\right)$ which in turn define the queue size distribution.

Theorem 3. *For any $\theta > 0$, there exists a unique equilibrium satisfying Definition 2.*

Theorem 3 establishes the existence of an equilibrium for the most general scenario where customers have to allocate their attention and acquire information about multiple aspects of the service environment. It is then also possible to compare equilibria that would emerge under different combinations of these uncertain dimensions. For example, it is possible to compare the equilibria when i) both service rate and queue length are uncertain, ii) only service rate is uncertain, and iii) only queue length is uncertain. Comparison of throughput under these scenarios would reveal insights as to whether it is better for the firm to provide information on service speed or queue length. We give a flavor of these insights in Figure 7. Note that “knowing more” (i.e., less uncertainty) does not necessarily lead to higher

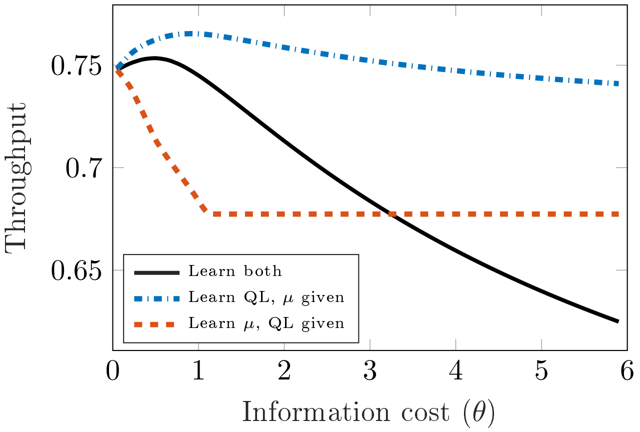


Figure 7 **Impact of learning service speed and/or queue length** ($R = 2.4, p = 0, C = 1, \lambda = 0.85, \Omega_\mu = \{0.5, 1, 1.5\}$)

throughput. We consistently find that throughput is higher when customers learn queue length instead of service speed, implying that providing visibility on service speed is more effective for the firm.

The extended model discussed in this section assumes that information cost is the same for each uncertain element of the environment the customer is learning about. In reality, however, some aspects of the service system may be easier to learn than others. For example, it is commonly argued that determining queue length is relatively easier than service speed. A rational customer should optimally allocate her attention taking into account the time-and-cost efficiency associated with learning about different aspects of the service system. The resulting choice behaviour would be more complex, but can be specified as a generalization of the GMNL choice (8), as shown in Hüttner et al. (2018). We can incorporate this choice function into our framework to characterize the equilibrium joining behavior with non-uniform information costs as well. Application of such models can produce more refined insights on optimal information provision strategies of the firm. Finally, we remark that our model can be enriched to include decisions beyond joining and balking. For example, it is possible that some customers may decide to "retry" joining the queue if they find it congested. Under mild conditions, it can be shown that an equilibrium exists under retrials. We omit further details in the interest of space.

8. Concluding Remarks

Limited attention is ubiquitous and learning is costly. In service systems, customers have to spend time and cognitive resources to determine queue lengths, estimate associated delays and translate this information into decisions. As information is costly, rational customers need to trade-off the benefits of information against its costs and have to make joining decisions based on partial information. Due to these information frictions, most queues are not fully visible or invisible, but rather opaque in operation. In this paper, we propose a tractable framework for opaque queues. At the core of our framework is rationally inattentive choice, which offers a micro-founded model of strategic customer behavior linking beliefs, rewards and information costs. Our framework covers the entire spectrum of queue opaqueness, naturally connecting the two well-known models of visible and invisible queues in the limits. Accordingly, we are able to provide a unified perspective and a comprehensive view on the effect of information cost (information prevalence) on throughput, revenue and social welfare.

We establish the existence and uniqueness of an equilibrium, and explore its sensitivity with respect to the underlying service characteristics. We identify perspicuous conditions for an opaque queue to be beneficial or detrimental to the firm from a throughput perspective, validate the robustness of these results when the service firm conducts price optimization, and also test social welfare implications. Instead of replicating these descriptive results here, it is perhaps better to focus on the managerial prescriptions they translate into, with some examples. Our results strongly suggest that firms should be most cautious about customer limited attention and their information provision strategies when there is a positive correlation between demand/congestion and how attractive the service is to the customers. This is because the effects of information frictions and hence information provision implications are reversed for high and low congested services.

When customers value a service highly and there is strong demand for it, service firms should intentionally leave some uncertainty around queue length, but not completely obstruct the information acquisition process. Disney, for example, is known to adopt such practices in theme parks via special layouts like serpentine lines that are partially blocked to disguise the length of the queue. Neither providing very clear queue length and delay information nor completely blocking queue visibility is productive, as they both lead to throughput losses. Service firms can take further advantage of customers' limited attention and presence of information costs by charging higher prices and increasing revenues in this case. If firm profits are relatively more significant compared to customer surplus, this might even be beneficial from a total social welfare perspective.

In contrast to above, for less congested firms offering a service that is not highly valued, opaque queues and partial hindrance of information acquisition is precisely what the firm should try to avoid. To the best of our knowledge, this has not been identified and noted in the extant literature. It is therefore in the interest of a low-congested public service office or drive-through fast-food restaurant to create a completely visible and transparent queuing system. It may even be better to completely obstruct the observation of the queue, but this may not be possible due to very nature of the process or physical constraints. It is optimal for a service firm operating in this regime to try to curb throughput losses by reducing prices, but we find that even that may not be sufficient to completely eradicate the losses due to information frictions.

We believe that our strategic queuing framework with rationally inattentive customers can serve as a useful instrument for service design. This is substantiated by the fact that the baseline model can be easily extended to accommodate finite waiting line capacity, multiple servers, as well as more complicated environments with multiple uncertain attributes where limited time and attention has to be allocated appropriately, among others. As we have shown with preliminary examples, these models can exploit trade-offs and provide valuable insights on the physical design attributes (number of servers, waiting room capacity) as well optimal provision of information about queue lengths, service speed, service quality, and possibly other salient characteristics of the service.

References

Ang E, Kwasnick S, Bayati M, Plambeck EL, Aratow M (2015) Accurate emergency department wait time prediction. *Manufacturing & Service Operations Management* 18(1):141–156.

Boyacı T, Akçay Y (2017) Pricing when customers have limited attention. *Management Science* 64(7):2973–3468.

Caplin A, Dean M, Leahy J (2016) Rational inattention, optimal consideration sets and stochastic choice, NYU working paper.

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

Cover TM, Thomas JA (2012) *Elements of information theory* (John Wiley & Sons, Hoboken, NJ).

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

Debo LG, Parlour CA, Rajan U (2012) Signaling quality via queues. *Management Science* 58(5):876–891.

Dessein W, Galeotti A, Santos T (2016) Rational inattention and organizational focus. *American Economic Review* 106(6):1522–36.

Dobson G, Pinker EJ (2006) The value of sharing lead time information. *IIE Transactions* 38(3):171–183.

Economou A, Kanta S (2008) Optimal balking strategies and pricing for the single server markovian queue with compartmented waiting space. *Queueing Systems* 59(3-4):237.

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

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

Hassin R (2016) *Rational queueing* (CRC press, Boca Raton, FL).

Hassin R, Haviv M (2003) *To queue or not to queue: Equilibrium behavior in queueing systems* (Kluwer Academic Publishers, Boston, MA).

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

Hu M, Li Y, Wang J (2017) Efficient ignorance: Information heterogeneity in a queue. *Management Science* 64(6):2473–2972.

Huang L, Liu H (2007) Rational inattention and portfolio selection. *The Journal of Finance* 62(4):1999–2040.

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

Huang T, Chen YJ (2015) Service systems with experience-based anecdotal reasoning customers. *Production and Operations Management* 24(5):778–790.

Hüttner F, Boyacı T, Akçay Y (2018) Consumer choice under limited attention when alternatives have different information costs, ESMT working paper 16-04 (R2).

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

Kremer M, Debo L (2015) Inferring quality from wait time. *Management Science* 62(10):3023–3038.

Maćkowiak B, Matějka F, Wiederholt M (2018) Rational inattention: A disciplined behavioral model, CERGE-EI working paper.

Matějka F (2015a) Rationally inattentive seller: Sales and discrete pricing. *The Review of Economic Studies* 83(3):1125–1155.

Matějka F (2015b) Rigid pricing and rationally inattentive consumer. *Journal of Economic Theory* 158(Part B):656–678.

Matějka F, McKay A (2015) Rational inattention to discrete choices: A new foundation for the multinomial logit model. *American Economic Review* 105(1):272–98.

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

Ren H, Huang T, Arifoglu K (2018) Managing service systems with unknown quality and customer anecdotal reasoning. *Production and Operations Management* 27(6):1038–1051.

Sallee JM (2014) Rational inattention and energy efficiency. *The Journal of Law and Economics* 57(3):781–820.

Simhon E, Hayel Y, Starobinski D, Zhu Q (2016) Optimal information disclosure policies in strategic queueing games. *Operations Research Letters* 44(1):109–113.

Sims CA (2003) Implications of rational inattention. *Journal of monetary Economics* 50(3):665–690.

Sims CA (2006) Rational inattention: Beyond the linear-quadratic case. *American Economic Review* 96(2):158–163.

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

Xu J, Hajek B (2013) The supermarket game. *Stochastic Systems* 3(2):405–441.

Appendix

Proof of Lemma 1 Note that (10) and (11) are from a direct application of Proposition 1 in Caplin et al. (2016) which provides necessary and sufficient conditions for a general discrete choice problem. Plugging $\overline{\pi} = 0$ in (10) and $\overline{\pi} = 1$ in (11), we obtain (12) and (13), respectively.

Proof of Theorem 1 Let us first show that the equilibrium in Definition 1 is stable. Assume that $\widetilde{\pi} < 1$ and $\theta \in [0, \infty)$. Then note that the queue distribution in steady-state is well-defined as the series

$$\sum_{k=1}^{\infty} \frac{\lambda_0 \lambda_1 \ldots \lambda_{k-1}}{\mu^k} = \sum_{k=1}^{\infty} a_k$$

converges. To see this let us use the ratio test. Note that

$$\lim_{k\to\infty} \left| \frac{a_{k+1}}{a_k} \right| = \lim_{k\to\infty} \left| \frac{\lambda}{\mu} \widetilde{\pi}_k \right| = \lim_{k\to\infty} \left| \frac{\lambda}{\mu} \frac{\widetilde{\pi} e^{v_k/\theta}}{1 - \widetilde{\pi} + \widetilde{\pi} e^{v_k/\theta}} \right| = 0$$

as $v_k = R - p - ck/\mu$ approaches to minus infinity as $k \to \infty$. This is true for any finite $\lambda > 0$. Note that when $\lambda \geq \mu$, equilibrium joining probability can not be 1 as expected waiting time for a customer is infinite. When $\theta = \infty$, the series is convergent if $\lambda < \mu$. Now we prove the existence and uniqueness.
Existence: Let us define

$$\pi_n(q) = \frac{q e^{v_n/\theta}}{1 - q + q e^{v_n/\theta}}$$

$$g_0(q) = \left(1 + \sum_{k=1}^{\infty} \rho^k \pi_0(q) \pi_1(q) \ldots \pi_{k-1}(q) \right)^{-1},$$

$$g_n(q) = g_0(q) \rho^n \pi_0(q) \pi_1(q) \ldots \pi_{n-1}(q) \quad \text{for } n \geq 1. \tag{21}$$

Our aim is to show that there exists at least a point satisfying $\overline{\pi}(q) = q$ where

$$\overline{\pi}(q) = \sum_n \pi_n(q) g_n(q). \tag{22}$$

Let $h(q) = \overline{\pi}(q) - q$. Note that $\pi_n(0) = 0$ for all $n \geq 0$ and consequently $g_0(0) = 1$ and $g_n(0) = 0$ for all $n \geq 1$. Using this and (22), it is clear that unconditional joining probability satisfies

$$\overline{\pi}(0) = \sum_n \frac{\overline{\pi}(0) e^{v_n/\theta}}{1 - \overline{\pi}(0) + \overline{\pi}(0) e^{v_n/\theta}} g_n(0) = \frac{\overline{\pi}(0) e^{v_0/\theta}}{1 - \overline{\pi}(0) + \overline{\pi}(0) e^{v_0/\theta}}$$

whose solution gives $\overline{\pi}(0) = 1$. This results in $h(0) = \overline{\pi}(0) = 1 > 0$, which is strictly positive. On the other hand, $h(1) = \overline{\pi}(1) - 1 \leq 0$ since $\overline{\pi}(1) \in [0,1]$. Since $h(q)$ is continuous in $[0,1]$, by the intermediate value theorem, there exists at least one point $q \in [0,1]$ satisfying $h(q) = 0$.

Uniqueness: To show uniqueness, it is enough to show that $h(q)$ is decreasing in q in the interval $[0,1]$. By (9), the unconditional joining probability is the solution to the following problem;

$$\overline{\pi}(q) := \underset{\pi \in [0,1]}{\arg\max} f(\pi; q) \quad \text{where} \tag{23}$$

$$f(\pi; q) = \theta \sum_n g_n(q) \log\left(\pi e^{v_n/\theta} + 1 - \pi\right). \tag{24}$$

For given $q \in [0,1]$, f is concave in π. As shown in Matějka and McKay (2015), there always exists a solution to (23) and if the vectors $e^{v_n/\theta}$ are linearly independent the solution is unique, which is exactly our case since v_n is strictly monotone in n. Taking partial derivative of f with respect to π gives

$$\frac{\partial}{\partial \pi} f(\pi; q) = \theta \sum_n g_n(q) \varphi_n(\pi; \theta) \quad \text{where}$$

$$\varphi_n(\pi; \theta) = \frac{e^{v_n/\theta} - 1}{\pi e^{v_n/\theta} + 1 - \pi}. \tag{25}$$

Note that $\varphi_n(\pi; \theta)$ is decreasing in n. [2] Now assume that q is increased to q'. By Lemma 2 below, it is clear that there exists a threshold $n^* \geq 0$ such that for all $n \leq n^*$, $g_n(q) - \varepsilon_n = g_n(q')$ and for all $n > n^*$, $g_n(q') = g_n(q) + \varepsilon_n$, for some sequence of $\varepsilon_n > 0$ with

$$\sum_{n=0}^{n^*} \varepsilon_n - \sum_{n=n^*+1}^{\infty} \varepsilon_n = 0. \tag{26}$$

Rewriting the first order derivative,

$$\begin{aligned}
\frac{\partial}{\partial \pi} f(\pi; q') &= \theta \sum_{n=0}^{n^*} g_n(q') \varphi_n(\pi; \theta) + \theta \sum_{n=n^*+1}^{\infty} g_n(q') \varphi_n(\pi; \theta) \\
&= \theta \sum_{n=0}^{n^*} (g_n(q) - \varepsilon_n) \varphi_n(\pi; \theta) + \theta \sum_{n=n^*+1}^{\infty} (g_n(q) + \varepsilon_n) \varphi_n(\pi; \theta) \\
&= \frac{\partial}{\partial \pi} f(\pi; q) - \theta \sum_{n=0}^{n^*} \varepsilon_n \varphi_n(\pi; \theta) + \theta \sum_{n=n^*+1}^{\infty} \varepsilon_n \varphi_n(\pi; \theta) \leq \frac{\partial}{\partial \pi} f(\pi; q)
\end{aligned}$$

since $\varphi_n(\pi; \theta)$ is decreasing in n. Then the value $\overline{\pi}(q)$ that satisfies the first order condition (i.e., the maximizer) is decreasing in q. Therefore, there is a unique $q^* \in [0,1]$ that satisfies $\overline{\pi}(q^*) = q^*$. $\square$

[2] We can rewrite $\varphi_n(\pi; \theta) = \frac{1}{\pi + \frac{1-\pi}{e^{v_n/\theta}}} - \frac{1}{\pi e^{v_n/\theta} + 1 - \pi}$. Observe that the first term is decreasing in n, whereas the second term is increasing in n.

Lemma 2. *For any $0 \le q_1 < q_2 \le 1$, $g_n(q_2) - g_n(q_1)$ is first negative, then positive as n increases.*

Proof Note that for $n = 0$, $g_0(q_2) - g_0(q_1) < 0$ since $g_0(q)$ is decreasing in q. We now consider two cases. Assume that $g_n(q_2) = g_n(q_1) + \varepsilon_n^g$ for some $n \ge 1$ where $\varepsilon_n^g > 0$. Similarly, assume that $\pi_n(q_2) = \pi_n(q_1) + \varepsilon_n^\pi$ for some $\varepsilon_n^\pi > 0$. Then, using (21), we show that

$$\begin{aligned} g_{n+1}(q_2) - g_{n+1}(q_1) &= \rho\pi_n(q_2)g_n(q_2) - \rho\pi_n(q_1)g_n(q_1) = \rho\pi_n(q_2)\left(g_n(q_1) + \varepsilon_n^g\right) - \rho\left(\pi_n(q_2) - \varepsilon_n^\pi\right)g_n(q_1) \\ &= \rho\pi_n(q_2)g_n(q_1) + \rho\pi_n(q_2)\varepsilon_n^g - \rho\pi_n(q_2)g_n(q_1) + \varepsilon_n^\pi\rho g_n(q_1) \\ &= \varepsilon_n^g\rho\pi_n(q_2) + \varepsilon_n^\pi\rho g_n(q_1) > 0. \end{aligned} \tag{27}$$

Now assume that $g_{n+1}(q_2) + \varepsilon_{n+1}^g = g_{n+1}(q_1)$ for some $n \ge 0$ where $\varepsilon_{n+1}^g > 0$. Assume to the contrary that $g_n(q_2) = g_n(q_1) + \varepsilon_n^g$ with $\varepsilon_n^g > 0$. Using (27) with $g_n(q_2) = g_n(q_1) + \varepsilon_n^g$ and $\pi_n(q_2) = \pi_n(q_1) + \varepsilon_n^\pi$,

$$g_{n+1}(q_2) - g_{n+1}(q_1) = \varepsilon_n^g\rho\pi_n(q_2) + \varepsilon_n^\pi\rho g_n(q_1) > 0$$

which is a contradiction. □

Proof of Corollary 1

1. When $\theta = 0$, for a given prior distribution g on queue size, the optimal solution to the maximization problem in (6) is $\pi_n = 1$ for all $n \le n_e$ and $\pi_n = 0$ for all $n \ge n_e + 1$ as v_n is decreasing in n. This is a visible queue which results in a capacitated $M/M/1/n_e$ system which gives (16).

2. When $\theta = \infty$, (14) implies that in equilibrium $\widetilde{\pi}_n = \widetilde{\pi}$ for all $n \ge 0$ and by (6), the problem is the same as in that of invisible queues. Let $\widetilde{\lambda} = \lambda\widetilde{\pi}$ be the equilibrium effective arrival rate. Note that $\widetilde{\lambda} \le \lambda$. We look for a symmetric Nash equilibrium where customers maximize $E_{G(\widetilde{\lambda})}\left[R - p - cn/\mu\right] = R - p - cW_q(\widetilde{\lambda})$ with $W_q(\widetilde{\lambda}) = \frac{\widetilde{\lambda}}{\mu(\mu - \widetilde{\lambda})}$. Then, it is clear that if $R - p - cW_q(\lambda) \ge 0$, then everyone joins. If, on the other hand, $\lambda \ge \mu$, $W_q = \infty$ and no one joins. Otherwise, unique equilibrium mixed strategy $\widetilde{\pi}$ satisfies $R - p - cW_q(\widetilde{\lambda}) = 0$ which yields the characterization in (17).

Proof of Proposition 1 The effect of R, p and c are unilateral since they only affect the utility. Therefore, we prove here that equilibrium joining fraction increases in utility, in general. Consider the optimization problem in (23). First we note that (25) is increasing in v_n for any π. Let us define the vector $v=[v_n;n\geq 0]$ and use the notation $v^{'}\geq v\equiv\left\{v_n^{'}\geq v_n;n\geq 0\right\}$. Using exactly the same arguments in Lemma 2, one can show that for any $v<v'$, $g_n(v')-g_n(v)$ is first negative, then positive as n increases. Then, it is clear that $\frac{\partial}{\partial\pi}f\left(\pi;q,v\right)$ is increasing in v, which implies that the maximizer defined in (23) will be increasing in v. Service rate μ, on the other hand, affect both utility values and congestion in the queue, i.e., queue size distribution. We proceed with a contradictory argument and utilize similar arguments in the proof of uniqueness of equilibrium. Let us denote $\widetilde{\pi}\left(\mu\right)$ as the equilibrium joining fraction as a function of service rate μ. Let $\mu_1<\mu_2$, and assume that $\widetilde{\pi}\left(\mu_1\right)>\widetilde{\pi}\left(\mu_2\right)$. Since $\widetilde{\pi}$ is a fixed point of $\left(23\right)$, $\overline{\pi}\left(\widetilde{\pi}\left(\mu_1\right)\right)>\overline{\pi}\left(\widetilde{\pi}\left(\mu_2\right)\right)$ where

$$\overline{\pi}\left(\widetilde{\pi}\left(\mu_i\right)\right)=\left\{\pi\in[0,1]:\theta\sum_n g_n\left(\widetilde{\pi}\left(\mu_i\right)\right)\varphi_n\left(\pi;\theta,\mu_i\right)=0\right\},\ \text{for } i\in\{1,2\}\quad\text{and}$$

$$\varphi_n\left(\pi;\theta,\mu\right)=\frac{e^{(R-p-cn/\mu)/\theta}-1}{\pi e^{(R-p-cn/\mu)/\theta}+1-\pi}.$$

First note that $\varphi_n\left(\pi;\theta,\mu_1\right)<\varphi_n\left(\pi;\theta,\mu_2\right)$. Furthermore, by Lemma 2, we know that $g_n(\widetilde{\pi}\left(\mu_1\right))-g_n(\widetilde{\pi}\left(\mu_2\right))$ is first negative, then positive as n increases, i.e., there exists a threshold $n^*\geq 0$ such that for all $n\leq n^*$, $g_n(\widetilde{\pi}\left(\mu_2\right))=g_n(\widetilde{\pi}\left(\mu_1\right))+\varepsilon_n$ and for all $n>n^*$, $g_n(\widetilde{\pi}\left(\mu_2\right))=g_n(\widetilde{\pi}\left(\mu_1\right))-\varepsilon_n$, for some sequence of $\varepsilon_n>0$ that satisfy $\left(26\right)$. Rewriting the first order derivative,

$$\begin{aligned}\frac{\partial}{\partial\pi}f\left(\pi;\mu_2\right)&=\theta\sum_{n=0}^{n^*}g_n\left(\widetilde{\pi}\left(\mu_2\right)\right)\varphi_n\left(\pi;\theta,\mu_2\right)+\theta\sum_{n=n^*+1}^{\infty}g_n\left(\widetilde{\pi}\left(\mu_2\right)\right)\varphi_n\left(\pi;\theta,\mu_2\right)\\&=\theta\sum_{n=0}^{n^*}\left(g_n(\widetilde{\pi}\left(\mu_1\right))+\varepsilon_n\right)\varphi_n\left(\pi;\theta,\mu_2\right)+\theta\sum_{n=n^*+1}^{\infty}\left(g_n(\widetilde{\pi}\left(\mu_1\right))-\varepsilon_n\right)\varphi_n\left(\pi;\theta,\mu_2\right)\\&=\theta\sum_{n=0}^{\infty}g_n(\widetilde{\pi}\left(\mu_1\right)\varphi_n\left(\pi;\theta,\mu_2\right)+\theta\sum_{n=0}^{n^*}\varepsilon_n\varphi_n\left(\pi;\theta,\mu_2\right)-\theta\sum_{n=n^*+1}^{\infty}\varepsilon_n\varphi_n\left(\pi;\theta,\mu_2\right)\geq\frac{\partial}{\partial\pi}f\left(\pi;\mu_1\right)\end{aligned}$$

since $\varphi_n\left(\pi;\theta,\mu\right)$ is decreasing in n. But this means that $\widetilde{\pi}\left(\mu_2\right)>\widetilde{\pi}\left(\mu_1\right)$, which is a contradiction. □

Proof of Proposition 2 From the first order condition, we obtain

$$R_I'\left(p\right)=\lambda\left[\widetilde{\pi}\left(p\right)+p\widetilde{\pi}'\left(p\right)\right]=0\Rightarrow\widetilde{\pi}'\left(p\right)=-\frac{\widetilde{\pi}\left(p\right)}{p}\tag{28}$$

where $\widetilde{\pi}'\left(p\right)$ denotes first order derivative with respect to p. Recall that $\widetilde{\pi}$ is the point that satisfy the fixed point equation $\overline{\pi}\left(q\right)=q$ where $\overline{\pi}\left(q\right)$ is defined in (23). Let,

$$F\left(p,\widetilde{\pi}\left(p\right)\right)=\overline{\pi}\left(\widetilde{\pi}\left(p\right),p\right)-\widetilde{\pi}\left(p\right)=0\quad\text{where}$$

$$\overline{\pi}\left(q,p\right):=\left\{x\in[0,1]:\theta\sum_{n}g_{n}\left(q,p\right)\frac{e^{v_{n}(p)/\theta}-1}{xe^{v_{n}(p)/\theta}+1-x}=0\right\}$$

and the dependence of $g_n\left(q,p\right)$ and $v_n\left(p\right)$ to p is clear from their definitions. To show that $R_I\left(p\right)$ is unimodal, it is enough to show that $\frac{\partial}{\partial p}F\left(\widetilde{\pi}\left(p\right),p\right)=0$ has a unique solution at point p that satisfies $R_I'\left(p\right)=0$. Using the chain rule, we take the first order derivative of $F\left(p,\widetilde{\pi}\left(p\right)\right)$ and use (28) to obtain

$$\tfrac{\partial}{\partial p}F\left(\widetilde{\pi}\left(p\right),p\right)=\tfrac{\partial}{\partial\widetilde{\pi}(p)}F\left(\widetilde{\pi}\left(p\right),p\right)\widetilde{\pi}'\left(p\right)=\left(\tfrac{\partial}{\partial\widetilde{\pi}(p)}\overline{\pi}\left(\widetilde{\pi}\left(p\right),p\right)-1\right)\widetilde{\pi}'\left(p\right)=0\Rightarrow\tfrac{\partial}{\partial\widetilde{\pi}(p)}\overline{\pi}\left(\widetilde{\pi}\left(p\right),p\right)=1\qquad(29)$$

since $\widetilde{\pi}'\left(p\right)<0$ (i.e., $\widetilde{\pi}\left(p\right)$ is decreasing in p by Corollary 1). We also know that $\overline{\pi}\left(q,p\right)$ is decreasing in q for a given p, which we proved in uniqueness part of Theorem 1. Furthermore, $\overline{\pi}\left(0,p\right)=1$ which means that there is a unique point that satisfy $\left(29\right)$. □

Proof of Theorem 2 Using necessary and sufficient conditions in Lemma 1, for a given belief g_0 (steady-state probability of zero customers in the system), the solution to the rational inattention problem with $v_0=R-p>0$ and $v_1=-T$ is

$$\overline{\pi}=\begin{cases}0 & \text{if } \frac{g_0}{1-e^{v_1/\theta}}-\frac{1-g_0}{e^{v_0/\theta}-1}<0\\ 1 & \text{if } \frac{g_0}{1-e^{v_1/\theta}}-\frac{1-g_0}{e^{v_0/\theta}-1}>1\\ \frac{g_0}{1-e^{v_1/\theta}}-\frac{1-g_0}{e^{v_0/\theta}-1} & \text{otherwise.}\end{cases}\qquad(30)$$

Noting $\widetilde{g}_0=\left(1+\rho\widetilde{\pi}_0\right)^{-1}$, the first element in (20) is the solution to the fixed point equation

$$\frac{\frac{1}{1+\rho\frac{\overline{\pi}^{*}e^{v_0/\theta}}{1-\overline{\pi}^{*}+\overline{\pi}^{*}e^{v_0/\theta}}}}{1-e^{v_1/\theta}}-\frac{1-\frac{1}{1+\rho\frac{\overline{\pi}^{*}e^{v_0/\theta}}{1-\overline{\pi}^{*}+\overline{\pi}^{*}e^{v_0/\theta}}}}{e^{v_0/\theta}-1}=\overline{\pi}^{*}.$$

Note that this point can not be negative since v_0 is assumed positive.However, it can be greater than one and in this case the unique equilibrium point is in the boundary, i.e., $\overline{\pi}^{*}=1$. □

Proof of Proposition 3

(1) Let us rewrite the first element in $\widetilde{\pi}$ in (20) as

$$\widetilde{\pi}=\frac{\left(e^{(R-p)/\theta}-1\right)}{\left(e^{T/\theta}-1\right)\left(\rho e^{(R-p-T)/\theta}+e^{(R-p-T)/\theta}-e^{-T/\theta}\right)}.$$

When $R-p\geq T$, it is clear that $\left(e^{(R-p)/\theta}-1\right)/\left(e^{T/\theta}-1\right)$ is increasing in θ. Furthermore, the denominator is decreasing in θ, which makes $\widetilde{\pi}$ increasing. When $R-p<T$, on the other hand, $\widetilde{\pi}$ may not

be monotone, yet, $\widetilde{\pi}$ is maximum when $\theta = 0$. To show this, let us first denote $\widetilde{\pi}(\theta)$ as a function of information cost. Assume to the contrary that for some $\theta > 0$, $\widetilde{\pi}(0) < \widetilde{\pi}(\theta)$, i.e.

$$\frac{1}{1+\rho} < \frac{\left(e^{(R-p)/\theta}-1\right)}{\left(1-e^{-T/\theta}\right)\left(\rho e^{(R-p)/\theta}+e^{(R-p)/\theta}-1\right).}$$

After some manipulation, it reduces to

$$\rho < \frac{e^{(R-p-T)/\theta}-e^{-T/\theta}}{\left(1-e^{(R-p-T)/\theta}\right)}.$$

Note that for $R-p<T$, the right hand side is increasing in θ and in the limit

$$\lim_{\theta\to\infty}\frac{e^{(R-p-T)/\theta}-e^{-T/\theta}}{\left(1-e^{(R-p-T)/\theta}\right)} = \lim_{\theta\to\infty}\frac{(R-p-T)/\theta e^{(R-p-T)/\theta}+T/\theta e^{-T/\theta}}{-(R-p-T)/\theta e^{(R-p-T)/\theta}} = \frac{(R-p)}{T-(R-p)}.$$

However, $\widetilde{\pi}(0) > \widetilde{\pi}(\infty)$ when $\rho > (R-p)/(T-(R-p))$ which is a contradiction. □

(2) For any $\theta \geq 0$, throughput is $\lambda\widetilde{\pi}_0/(1+\rho\widetilde{\pi}_0) = \lambda/(1/\widetilde{\pi}_0+\rho)$. When $\theta = 0$, throughput is $\lambda/(1+\rho)$. Since $\widetilde{\pi}_0 \in [0,1]$, $\lambda/(1+\rho) \geq \lambda/(1/\widetilde{\pi}_0+\rho)$. □

Proof of Theorem 3 Let $\rho_j = \lambda/\mu_j$ and define

$$\overline{\pi}_n(q;i,j) = \frac{qe^{v_n(i,j)/\theta}}{1-q+qe^{v_n(i,j)/\theta}} \text{ for } i\in\{1,..,K_R\}, j\in\{1,..,K_\mu\}, n\geq 0 \quad \text{and}$$

$$\overline{\pi}_n(q;j) = \sum_{i=1}^{K_R}\overline{\pi}_n(q;i,j)\,\overline{g}_n(q;j)\,h_{R,\mu}(i,j) \text{ for } n\geq 0 \text{ and } j\in\{1,..,K_\mu\}$$

with queue size distribution

$$\overline{g}_0(q;j) = \frac{1}{1+\sum_{n=1}^{\infty}\rho_j^n\overline{\pi}_0(q;j)\,\overline{\pi}_1(q;j)\ldots\overline{\pi}_{n-1}(q;j)},$$

$$\overline{g}_n(q;j) = \overline{g}_0(q;j)\,\rho_j^n\overline{\pi}_0(q;j)\,\overline{\pi}_1(q;j)\ldots\overline{\pi}_{n-1}(q;j) \quad \text{for } n\geq 1.$$

Consider the following maximization problem:

$$\overline{\pi}(q) = \arg\max_{\pi\in[0,1]}\left[f(\pi;q) = \theta\sum_{n=0}^{\infty}\sum_{i=1}^{K_R}\sum_{k=1}^{K_\mu}\overline{g}_n(q;j)\,h_{R,\mu}(i,j)\,log\left(\pi e^{v_n(i,j)/\theta}+1-\pi\right)\right] \tag{31}$$

Let $h(q) = \overline{\pi}(q) - q$. Note that $h(0) = \overline{\pi}(0) \geq 0$ and $h(1) = \overline{\pi}(1) - 1 \leq 0$ since $\overline{\pi}(q) \in [0,1]$. Since $h(q)$ is continuous in $[0,1]$, by the intermediate value theorem, there exists at least one point $q \in [0,1]$ satisfying $h(q) = 0$.

Uniqueness: For a given $q \in [0,1]$, the objective function in (31) is concave in π. We can write the first order derivative of the objective in (31) with respect to π as

$$\frac{\partial}{\partial \pi} f(\pi; q) = \theta \sum_{i=1}^{K_R} \sum_{k=1}^{K_\mu} h_{R,\mu}(i,j) \sum_{n=0}^{\infty} \overline{g}_n(q;j)\, \varphi_n(\pi;\theta) \quad \text{where}$$

$$\varphi_{n,k}(\pi;\theta) = \frac{e^{v_n(i,j)/\theta} - 1}{\pi e^{v_n(i,j)/\theta} + 1 - \pi}. \tag{32}$$

Observe that for fixed i, j, $\varphi_{n,k}(\pi;\theta)$ is decreasing in n and by Lemma 2 and by the same arguments in the proof of Theorem 1, $\sum_{n=0}^{\infty} \overline{g}_n(q;j)\, \varphi_n(\pi;\theta)$ is decreasing in π. Note that $\frac{\partial}{\partial \pi} f(\pi;q)$ is just expectation of this value and hence it also decreases in π. The same arguments follow as in Theorem 1 and there exists a unique $q^* \in [0,1]$ that satisfies $\overline{\pi}(q^*) = q^*$. □

Stability of the Equilibrium

In this section, we show how the equilibrium in Definition 1 can be attained in an adaptive way. We use time periods indexed as $t \in \{0,1,2,...\}$ and assume that each period is long enough for the system to reach steady-state. At $t = 1$, assume that customers start with an arbitrary belief about percentage of joining customers (market share) q_0 by which they form their prior belief $G(q_0)$ about queue size distribution which is defined in (21). Customers' best response in period $t = 1$ given $G(q_0)$ is then $q_1 = \overline{\pi}(q_0)$ which is defined in (23). At any period $t \geq 1$, customers use the average market share to form their prior belief about queue size distribution. More specifically, their prior belief is $G(\overline{q}_{t-1})$ where $\overline{q}_{t-1} = \sum_{k=0}^{t-1} q_k/t$. Then, the resulting unconditional joining probability is $q_t = \overline{\pi}(\overline{q}_{t-1})$. In the following proposition, we show that customer behaviour $\overline{q}_t$ converges to the equilibrium joining probability q^* that satisfies $\overline{\pi}(q^*) = q^*$, i.e., equilibrium given in Definition 1. However, before that we give the following useful lemma that connects optimal rational inattentive behaviour to the prior distribution which is a slightly modified version of the Lemma 16.1.1 in Cover and Thomas (2012) for log-optimal portfolios.

Lemma 3. *$\overline{\pi}(q)$ is convex in q.*

Proof First note that $f(\pi;q)$ in (24) is linear in distribution, i.e., $G(q)$. Let $G(q_1)$ and $G(q_2)$ be two distributions with corresponding joining probabilities $\overline{\pi}(q_1)$ and $\overline{\pi}(q_2)$. From linearity,

$$\overline{\pi}(\lambda q_1 + (1-\lambda) q_2) = f(\overline{\pi}(\lambda q_1 + (1-\lambda) q_2), \lambda q_1 + (1-\lambda) q_2)$$

$$
\begin{aligned}
&= \lambda f\left(\overline{\pi}\left(\lambda q_1 + (1-\lambda)\, q_2\right), q_1\right) + (1-\lambda)\, f\left(\overline{\pi}\left(\lambda q_1 + (1-\lambda)\, q_2\right), q_2\right) \\
&\leq \lambda f\left(\overline{\pi}\left(q_1\right), q_1\right) + (1-\lambda)\, f\left(\overline{\pi}\left(q_2\right), q_2\right)
\end{aligned}
$$

where the last inequality is due to optimality. □

PROPOSITION 4. *The sequence* $q = \{q_t; t \geq 0\}$ *converges to* q^* *which satisfies* $\overline{\pi}\left(q^*\right) = q^*$.

Proof Using Lemma 3, we can write

$$
\begin{aligned}
q_{t+1} = \overline{\pi}\left(\sum_{k=0}^{t} \frac{q_k}{t}\right) &= \overline{\pi}\left(\left(\frac{t}{t+1}\right)\sum_{k=0}^{t-1}\frac{q_k}{t} + \left(\frac{1}{t+1}\right) q_t\right) \\
&\leq \left(\frac{t}{t+1}\right)\overline{\pi}\left(\sum_{k=0}^{t-1}\frac{q_k}{t}\right) + \left(\frac{1}{t+1}\right)\overline{\pi}\left(q_t\right)\left(\frac{t}{t+1}\right) q_t + \left(\frac{1}{t+1}\right)\overline{\pi}\left(q_t\right).
\end{aligned}
$$

Then, note that

$$
q_{t+1} - q_t \leq \frac{\overline{\pi}\left(q_t\right) - q_t}{t+1}.
$$

Since for any t, q_t and $\overline{\pi}\left(q_t\right)$ are in $[0,1]$, as $t \to \infty$, $q_{t+1} - q_t \to 0$. That is, the resulting joining probabilities converge to the same number, i.e., $\underset{t\to\infty}{lim}\ q_t = q^*$. It is clear that, $\overline{\pi}\left(q^*\right) = q^*$. □

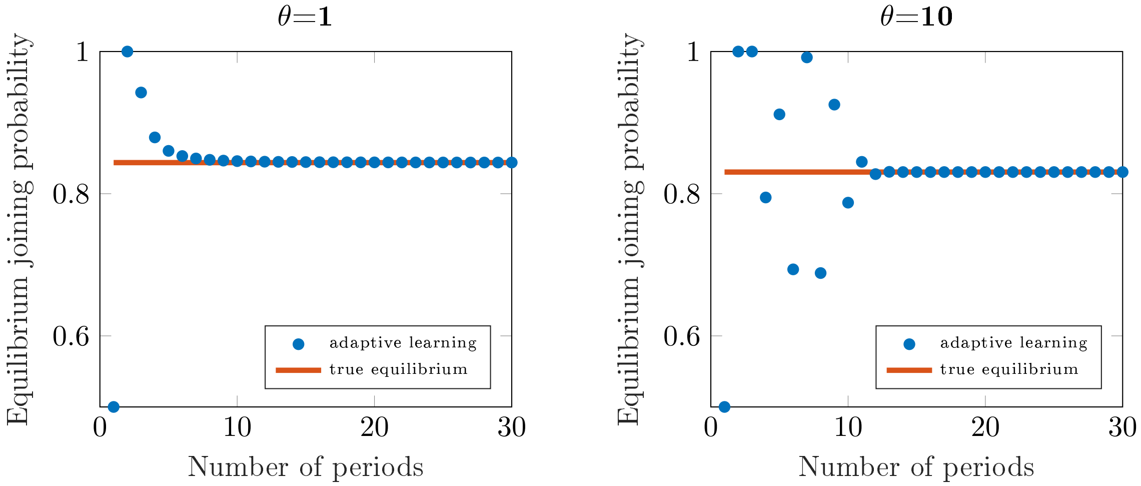


Figure 8 **Convergence to the system equilibrium** $(R = 2.8, p = 0, c = 1, \mu = 1, \lambda = 0.9)$

In Figure 8, we provide an illustrative example on convergence to the equilibrium for two different information cost values. We arbitrarily assume that customers at period $t = 1$ start constructing their beliefs using $q_0 = 0.5$. The horizontal line represents the true equilibrium value in these figures. Note that customers construct the true belief and hence equilibrium is reached very quickly. As θ gets higher, convergence speed gets slightly slower.

Recent ESMT Working Papers

	ESMT No.
The Coleman-Shapley-index: Being decisive within the coalition of the interested André Casajus, HHL Leipzig Graduate School of Management Frank Huettner, ESMT Berlin	18-03
Reverse privatization as a reaction to the competitive environment: Evidence from solid waste collection in Germany Juri Demuth, E.CA Economics Hans W. Friederiszick, ESMT European School of Management and Technology and E.CA Economics Steffen Reinhold**, E.CA Economics**	18-02
Knowing me, knowing you: Inventor mobility and the formation of technology-oriented alliances Stefan Wagner, ESMT Berlin Martin C. Goossen, Tilburg University	18-01
Static or dynamic efficiency: Horizontal merger effects in the wireless telecommunications industry Michał Grajek, ESMT European School of Management and Technology Klaus Gugler, Vienna University of Economics and Business Tobias Kretschmer, Ludwig Maximilian University of Munich **Ion Mişcişin, University of Vienna**	17-04
Brand positioning and consumer taste information Arcan Nalca, Smith School of Business, Queen's University Tamer Boyaci, ESMT European School of Management and Technology Saibal Ray, Desautels Faculty of Management, McGill University	17-01 (R1)