Managing Service Systems Via Disguised Queues: The Role of Strategic Customer Behavior

Eren B. Çil
Lundquist College of Business, University of Oregon, Eugene, Oregon, erencil@uoregon.edu

Touristic attractions, such as observatory decks, boat tours, museums, have recently started to manage their operations via hiding or obstructing some parts of their waiting lines, which we will refer to as *queue disguising* behavior. If customers are not aware of these disguised parts of the queues, the firm can easily boost its revenues by engaging in queue disguising behavior. However, it is not obvious that the firm benefits from disguised queues when customers are strategic, i.e., when they anticipate that the firm hides some parts of the waiting lines. Our goal in this paper is to investigate the impacts of strategic customer behavior on the firm's queue disguising decisions and profits. To this end, we consider a firm that can hide the initial parts of the waiting lines that form in front of its service center, and thus customers have to make their service requests based on the observable portion of the queue. We also suppose that some of the customers, which will be referred to as *strategic customers*, penalize the firm by not revisiting the firm in the future if they are fooled to request service because of the disguised queues. In the absence of strategic customers, we find that the firm's optimal queue disguising strategy yields only negligible profit improvements, relative to letting customers observe the entire waiting line. As a matter of fact, these minimal profit improvements vanish, and the queue disguising becomes harmful as the cost of holding customers in the queue increases. One, then, may expect that the queue disguising should not be a profitable strategy in a system with strategic customers because the firm has to account for the negative implications of disguised queues on customers' behavior. We establish that this intuition is only true when the firm optimally hides some of its waiting spots while facing non-strategic customer. If the queue disguising is already harmful under non-strategic customers, we surprisingly find that the firm starts to obtain sizable benefits from queue disguising as the customers become more strategic. More interestingly, the firm's profit gains from disguised queues increase as the holding cost increases in contrast to the case without strategic customers. This result brings to light a crucial insight for the service environments we consider in this paper as these systems bear non-negligible costs to keep the waiting customers happy: The firms can significantly benefit from disguised queues if customers act strategically whereas any attempts to obstruct the waiting lines hurt a firm facing non-strategic customers.

Key words: Strategic customers; queue management; finite queues; asymptotic analysis.

1. Introduction

Waiting lines, or queues as the British would say, are often inevitable in service environments. Ranging from emergency rooms to ball park entrances, patrons frequently end up waiting for some time before getting the attention of a service provider. Waiting is not only a concern for people visiting the service facilities. As waiting can be associated with numerous behavioral, psychological, and micro-economical interpretations, service providers also struggle with the unwelcome consequences of queues. A "one-fit-all" solution never exists to handle the customers' waiting experience in service systems. Depending on the type of the service encounter, customer expectations, and physical conditions of the service facility, the "right" solution to mitigate the undesirable implications of customer waiting vary significantly. Banks, government offices (e.g., DMV), and pharmacies let customers enter an "invisible" queue, so that customers do not need to physically stand in the queue. In services where physical presence in queue is required, such as toll booths, security check points, and venue entrances, service providers try to make the waiting more tolerable by allowing customers to choose the queue they want to join. As mentioned in Fitzsimmons et al. (2014), other efforts to ease the discomfort of staying in the queue include setting up a more pleasant waiting area, communicating the anticipated waiting times with customers, and providing "service-related diversions" such as handing out dinner menus. Service firms can also manage their waiting lines by following a priority rule. Some services prioritize customers according to their observable characteristic while in some others, customers can purchase the service priority. All of these different practices affect the dynamics and the perception of waiting and may help the firm better serve its customers if used in the right context.

As another tactic to manage its waiting lines, a service firm may intentionally obstruct how customers observe the waiting area. Specifically, firms can hide the initial parts of the waiting lines, so that customers see a shorter version of the queue. We will refer to such a practice as *queue disguising* strategy. The queue disguising strategy has, recently, become a widespread exercise in touristic attractions, such as theme parks, observatory decks, museums, boat tours, etc. For instance, one has to go through several unexpected hidden waiting rooms before taking the last elevator to the *Observation Deck* at the Empire State Building. Patrons of theme parks also experience that the "real" queue usually turns out to be longer than the queue they initially see.

The queue disguising strategy, most probably, has become popular because it can help a service firm to manipulate the customer demand and thus may improve the firm's profits, especially if the customers are not aware of the disguised queues. However, it is not obvious that firms can benefit from obstructing its waiting lines when customers are strategic, i.e., when customers react and change their behavior once they realize the existence of the disguised queues. Hence, in this paper, we aim at investigating the impacts of strategic customer behavior on the efficiency of the queue disguising strategy as a demand management tool. To this end, we consider a firm that can hide the initial parts of the waiting line that forms in front of its service center, and thus customers have to make their service requests based on the observable portion of the queue. In particular, customers request service and join the queue if the number of customers in the observable part of the waiting line is below a threshold which depends on the speed of the service and their tolerance/patience for waiting.

When the firm hides a portion of its waiting line, some of the customers join the queue despite that fact that doing so is sub-optimal had they known the existence of the hidden portion of the queue. If these customers realize that the firm fooled them to join the queue, then they may feel a dissatisfaction. It is also natural to expect that some of these customers will penalize the firm if the levels of their dissatisfaction are so high. Specifically, we suppose that some of the customers, which will be referred to as *strategic customers*, react to the existence of disguised queues by not revisiting the firm in the future if they fallaciously make a service request. We capture the strategic customers' future punishment by studying the firm's queue disguising problem as a two-period model. Our two-period model is aligned with how most of the service environments we focus on this paper function because customers revisit these services multiple times, at least at a seasonal level. In fact, in theme parks, customers typically take the same ride several times on the same day. It is worth noting that our model can also apply to the service settings without customer revisits as long as the future customers have access to (mostly digital) platforms where the past customers can share their experience, such as online travel forums and blogs.

According to the 2015 report from *Themed Entertainment Association* (See Rubin (2016)), the top 25 amusement/theme parks attracted 235 million visits worldwide, which was a 5.4% increase from 2014. Each of the top three parks from this list had more than 15 million visitors. The same report also states that the attendance for the top five museums and water parks were 36 and 10 million in 2015. All of these astonishing numbers clearly suggest that the service systems that this paper review serve large amount of people in a short time. To capture this feature, we consider the firm's service center as a large-scale service system, where the service and arrival rates are high, in this paper.

As the main objective of this paper is to analyze how strategic customer behavior affects the firm's queue disguising decisions, we first focus on the case where customers are not strategic. When the firm faces non-strategic customers, we show that the firm finds queue disguising optimal only when the customers are very impatient and the cost of keeping the customers in the queue, which will be referred to as *holding cost*, is very small. As the customers become more tolerant for waiting or the holding cost increases, we show that the firm hides a smaller portion of its waiting lines. In fact, once the customer patience or the holding cost exceeds a critical level, the firm completely abandons queue disguising. We also examine the extent to which disguised queues increase profit, relative to letting customers observe the entire waiting line. We find that the firm's profit gains from the disguised queues are capped by a threshold which points to negligible profit improvements in large-scale systems.

We next study the case where customers are strategic. When customers are strategic, the firm has to account for the negative implications of disguised queues on customers' behavior. One, then, may expect that the queue disguising should not be a profitable strategy in a system with strategic customers because the firm's benefits from queue disguising is already insignificant in the absence of strategic customers. We establish that this intuition is only partially true. In the case where the firm optimally hides some of its waiting spots while facing non-strategic customer, we find that the strategic customer behavior dampens the profit gains queue disguising strategy yields for the firm. However, if the queue disguising is already harmful under non-strategic customers, we surprisingly find that the firm starts to obtain sizable benefits from employing disguised queues as the customers become more strategic. Along the same lines with these findings, the structural properties of the firm's optimal queue disguising strategy also take a U-turn as a response to the presence of strategic customers. In particular, the firm hides more spots in its waiting line as the customers' patience or the holding cost increases, which completely contrasts with the firm's optimal decisions when facing non-strategic customers. Our analysis underscores the importance of taking customer behavior into account in managing service systems.

On the theoretical front, we derive approximations for the key performance metrics of the firm's service center by studying the asymptotic behavior of the original system. Specifically, we consider a sequence of systems that are the replicas of the firm's service center in a parametric regime where the demand and the arrival rates grow unboundedly. As the service facilities we consider in this paper tend to process customers at a fast pace and attract high volumes of customer demand, our asymptotic analysis lead to efficient and accurate approximations for many crucial system metrics such as server utilization and average queue length.

The rest of the paper is organized as follows. We survey the related literature in Section 2. Then, in Section 3, we describe the basics of our model. Section 4 builds an approximation for the firm's profit functions based on the limiting behavior of the original system in a parametric regime where the demand and the arrival rates are high. We analyze the firm's queue disguising problem both in the absence and the presence of the strategic customers in Section 5. Section 6 concludes the paper.

2. Literature Review

The previous work related with our paper can be divided into two categories. The first category consists of research that studies the applications of strategic customers in service systems. The second line of research focuses on sharing system information, such as waiting times and queue lengths, with customers in service business.

The decision making process of price and time sensitive customers in service systems have attracted the attention of researchers for many years. The literature on this research stream dates back to Naor's seminal work (See Naor (1969)), which analyze customer behavior in a single-server queueing system. In this paper, we build our model based on the Naor's model, where customers can see the entire waiting line upon their arrival and request service based on the queue length they observe. Our major deviation from Naor (1969) is that we allow the firm to partially and intentionally hide the waiting line that forms in front of its service center. Furthermore, we consider a two-period model where some of the customers may react to the firm's queue disguising strategy in the second period. Motivated by Naor (1969), many researchers study the pricing problem of a monopoly facing price- and delay-sensitive customers in various settings (See De Vany (1976), Mendelson and Whang (1990), Afeche and Mendelson (2004), Boudali and Economou (2012)). Another body of research that is motivated by Naor (1969) considers the competition among service providers who make pricing and/or service capacity decisions. We refer the reader to Hassin and Haviv (2003) for an extensive summary of the early attempts to model price and service competition. More recent examples service competition models with a focus on customers' demand decisions are Cachon and Zhang (2007), Allon and Federgruen (2008), Li et al. (2012), and Chiu et al. (2014).

The other growing body of literature that is related to our paper studies service systems where firms disclose information about their waiting lines and/or inventories. In one of the earliest attempt to explore the role of information sharing in service environments, Hassin (1986) shows that a firm may find it optimal not to reveal the queue length despite sharing more information being socially optimal. Another earlier paper in this line of research is Whitt (1999). Whitt (1999) compares two models of multi-server queues with limited buffer capacity where one model provides no information while the other one is sharing the queue length information. The paper shows that information sharing improves the throughput of the system but deteriorates the likelihood of serving customers without letting them wait. Guo and Zipkin (2007) also compares service systems varying in terms of the level of information provided. The paper considers three different information structures: (i) no information, (ii) queue- length information, and (iii) exact waiting-time information. Guo and Zipkin (2007) establishes that the information sharing may have a non-monotone impact on the system performance. The paper's findings do not completely echo Whitt (1999) because two papers employ different customer-choice mechanisms. Ziani et al. (2015) performs a similar comparison when service requests arrive in batches. More recently, Simhon et al. (2016) and Kim and Kim (2017) study the information disclosure decisions of a firm operating a single-server system. Interestingly, they show that revealing the queue length is not optimal when the queue is short. In addition to these papers that explores the queue length information sharing, Armony and Maglaras (2004) studies a call center model where the firm shares its system information via committing to a guaranteed time for the service start. Jouini et al. (2011) is another paper where the firm chooses to share information without revealing queue length. It considers delay announcement as the medium of information sharing. All of the above mentioned papers assumes that the firm accurately shares the queue or delay information once it decides to do so. In our paper, we model a service system where the firm deliberately provide an imprecise information to influence customers' decisions.

In the information sharing literature, our paper is closer to the papers that study service environments where partial and/or uninformative information is shared. Examples of these papers are Dobson and Pinker (2006), Economou and Kanta (2008), and Allon et al. (2011). The imprecise information sharing is a feasible option in the models reviewed in these papers because customers cannot fully observe the queues or the inventories. In our paper, customers can observe the entire waiting lines, except for the disguised portion. More importantly, the main focus of our paper is to study the implications of strategic customer behavior, which is not considered in any of these papers.

3. The Model

We study a two-period model in which a profit maximizing service firm caters to time sensitive customers. The firm owns a single service facility, and the service times are independent and exponentially distributed with mean $1/\mu_o$. In both periods, customer demand for the service is generated according to a Poisson process with rate λ_o. This forms the "*potential demand*" for the firm. Following the conventional terminology from the literature, we refer to the ratio λ_o/μ_o as the *system load* and denote it by ρ_o.

On the demand side, each customer incurs a waiting cost of c per unit time spent in the system and obtains a value of R after a successful completion of service. The firm earns a value of V when a customer is served and incurs a holding cost of h per customer per unit time until her service starts. The driver of the holding cost can be loss of goodwill, the opportunity cost associated with the customer not being able to generate revenues, and in some settings the actual cost of holding the customer. For instance, theme parks hire entertainment crews in order to alleviate the pain of standing in line for customers.

When customers arrive to the system, they decide whether to request service or not. Customers deciding not to request service obtain a utility of zero. If a customer requests service, she joins the queue in front of the server and waits for her service to commence. The service of a customer starts immediately if the server is available. Customers who waits in the queue are served in a *First-Come-First-Served* manner. While making their decisions, customers can observe the number of people in the queue to the extent that is allowed by the firm. To be specific, the firm can hide the first $K \geq 0$ spots in its waiting lines to influence the decisions of the customers. We refer to the number of disguised waiting spots, K, as the *disguised-queue size.* We suppose customers are unaware of the firm's queue disguising decision when they first visit the firm. The main objective of the firm is to find the optimal disguised-queue size K to maximize its profits.

In his seminal paper, Naor (1969) establishes that customers request service as long as they see $\lfloor R\mu_o/c \rfloor$ people in the system when the queue is fully observable, i.e., when the disguised-queue size K is zero. We will refer to $\lfloor R\mu_o/c \rfloor$ as the *Naor threshold* and denote it by τ_N. On the other hand, when the firm chooses a non-zero K, customers will behave similar to Naor's model but will keep requesting service until there are $K+\tau_N$ people in the system because they can only partially observe the waiting line. Hence, in the first period, the firm's service facility operates as an $M/M/1/K+\tau_N$ queue with arrival rate λ_o and service rate μ_o given the queue disguising decision of the firm. Then, we can write the firm's profit in the first period as

$$\Pi_1(K,\lambda_o,\mu_o) = V * \mu_o * \sigma(K,\lambda_o,\mu_o) + h * Q(K,\lambda_o,\mu_o), \tag{1}$$

where $\sigma(K,\lambda,\mu)$ and $Q(K,\lambda,\mu)$ are the average utilization of the server and the average queue length in an $M/M/1/K+\tau_N$ queue with arrival rate λ and service rate μ. It is worth noting that the above profit is the firm's profit rate per unit time, but throughout the paper we refer to it as the profit for ease of explanation.

When the customers request service up to the Naor threshold, they join the queue only if joining is in their self-interest. However, the queue disguising strategy of the firm may fool some customers to make a service request although such a decision is not in the best interest of themselves. As a result of this misleading circumstances, customers may be dissatisfied and decide to punish the firm. In particular, we suppose that $\gamma \geq 0$ fraction of customers who end up waiting for more than the Naor threshold will decide to leave the firm in the second period. Thus, letting $\beta(K, \lambda_o, \mu_o)$ be the probability with which the customers wait for more than or equal to τ_N people in the first period, the *effective arrival rate* in the second period will be $\lambda[1-\gamma\beta(K, \lambda_o, \mu_o)]$. As the customers staying with the firm (who form the effective demand) continue to request service until they find $K+\tau_N$ people in the system, the firm's service facility, in the second period, operates as an $M/M/1/K+\tau_N$ queue with arrival rate $\lambda[1-\gamma\beta(K, \lambda_o, \mu_o)]$ and service rate μ_o. Then, the firm's profit in the second period becomes

$$\Pi_2(K, \lambda_o, \mu_o) = \Pi_1(K, \lambda[1-\gamma\beta(K, \lambda_o, \mu_o)], \mu_o). \tag{2}$$

Notice that $\beta(K, \lambda, \mu) \equiv \sum_{i=\tau_N}^{K+\tau_N-1} P_i(K, \lambda, \mu)$, where $P_i(K, \lambda, \mu)$ is the steady state probability of having i customers in an $M/M/1/K+\tau_N$ queue with arrival rate λ and service rate μ.

Given the setup described above, the firm solves the following optimization problem to determine the optimal disguised-queue size:

$$\Pi^*(\lambda_o, \mu_o) \equiv \max_{K \geq 0} \Pi_1(K, \lambda_o, \mu_o) + \Pi_2(K, \lambda_o, \mu_o). \tag{3}$$

As it is evident from the above discussions, the customer reward R and waiting cost c affects the system only through the Naor threshold τ_N. Hence, we will henceforth treat τ_N as a system parameter replacing these two parameters. By doing so, we will use the Naor threshold as the measure of customers' tolerance/patience for waiting. Besides its notational convenience, considering τ_N as a parameter helps us to relate our analysis and findings to real life because it is easier to conceptualize the Naor thresholds of the customers than their valuations and waiting costs. To illustrate that, consider an instance of our model where the service rate is 200 customers per hour, customer reward is \$10 and waiting cost is \$20 per hour. It might be difficult to put this description of the example into perspective because the concepts of reward and waiting cost are hard to observe and measure. We believe it is more natural to describe the same example by noting that the Naor threshold is 100.

4. Approximations

In the previous section, we introduce the details of how we model the firm's service system and customers' behavior. According to the model we construct, once the customers make their service request decisions, the firm's operations in both periods simply run as a single server queueing system with finite waiting capacity. Although single server systems with finite queues have explicit formulas for the performance metrics such as server utilization and average queue length, these formulas do not result in tractable enough expressions to study and solve the firm's problem. More importantly, our main goal in this paper is not just to obtain optimal solutions. In fact, we aim at assessing the benefits of the queue disguising strategy and exploring the impacts of strategic customer behavior on these benefits. The solutions we can obtain studying the firm's original system, which are often quite complicated and implicit, makes our main objective analytically intractable. Hence, in this section, we shall approximate the firm's profit functions by studying the limiting behavior of the original system in a parametric regime where the demand and the arrival rates are high. As the service facilities we consider in this paper tend to process customers at a fast pace and attract high volumes of customer demand, the asymptotic analysis lead to promising and accurate approximations.

To construct our approximations, we consider a sequence of systems, indexed by n, that are the replicas of the firm's service facility, except for the arrival and the demand rate. We denote the service and the arrival rates in the n^{th} replica system by μ_n and λ_n, respectively. Furthermore, we denote the system load of the n^{th} replica by ρ_n. In addition to the large-scale nature of the service systems we study, they are typically heavily loaded systems. In other words, the system load is close to 1. Hence, while constructing our approximations, we consider a sequence of systems where

ρ_n converges to 1 as the size of the system grow. In particular, we let

$$\rho_n \equiv 1 - \omega/\mu_n,$$

where ω is a positive constant. We do not impose any structure on ω but the choice of the constant ω turns out to be crucial to obtain accurate approximations for the system dynamics of the firm's original service facility.

In the following subsections, we first study the firm's profit in the first period and then focus on the second period profit.

4.1. The First Period

In this subsection, we construct our approximation for the firm's profit function in the first period by studying the sequence of the replica systems described above. To this end, we first establish the asymptotic behaviors of the server utilization, $\sigma(K, \lambda, \mu)$, and the average queue length, $Q(K, \lambda, \mu)$, functions along this sequence as the system size grows in Proposition 1.

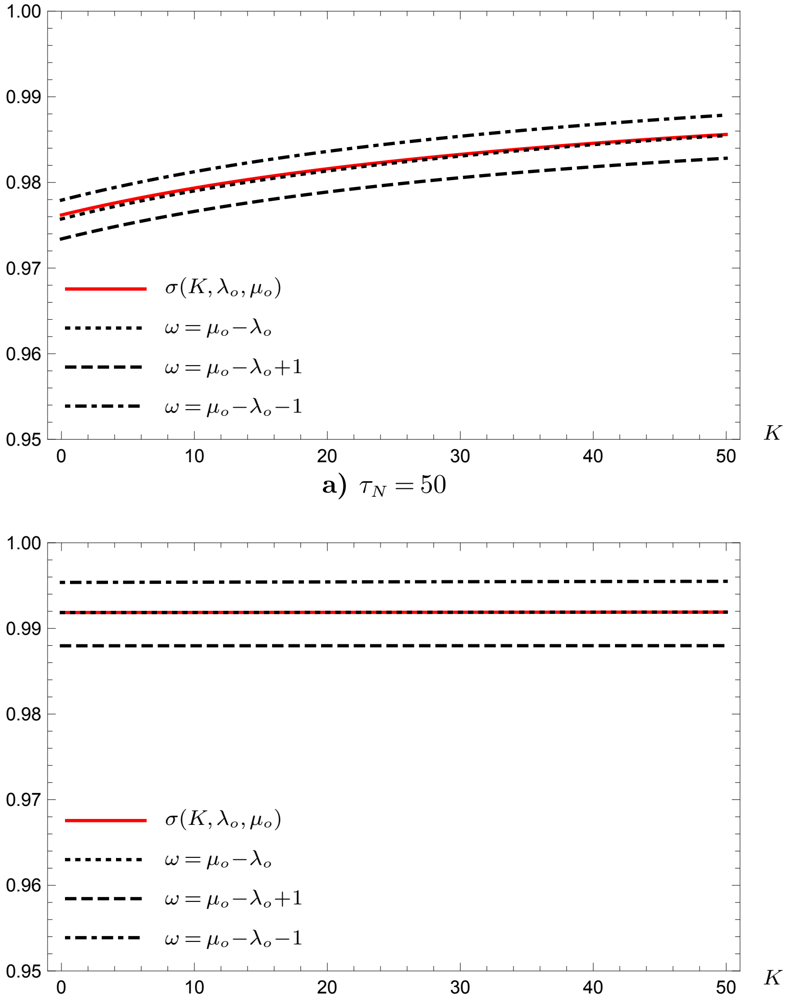
PROPOSITION 1. *Consider the sequence of systems introduced at the beginning of Section 4. If $\mu_n \to \infty$ as $n \to \infty$, then we have that*

$$\lim_{n\to\infty} \left[1 - \sigma(\kappa\mu_n, \lambda_n, \mu_n)\right]\mu_n = \omega\left(1 + \frac{1}{e^{(\kappa+\tau_N/\mu_o)\omega} - 1}\right), \text{ and} \tag{4}$$

$$\lim_{n\to\infty} \left[Q(\kappa\mu_n, \lambda_n, \mu_n)\right]/\mu_n = \frac{1}{\omega} - \frac{\kappa + \tau_N/\mu_o}{e^{(\kappa+\tau_N/\mu_o)\omega} - 1}. \tag{5}$$

Proposition 1 proves the asymptotic behaviors of the sever idleness and the average queue length. To derive the above results, we let the firm's queue disguising decision be a fraction of the service rate along the sequence of the replica systems. We denote that fraction by κ. By doing so, we avoid a limiting result that is independent of the firm's choice of the disguised-queue size as the system size grows.

We can use the results in Proposition 1 to formulate approximations for the firm's original system in the first period. For instance, we can approximate the server utilization using the limit result from (4), after dividing it by the service rate. As we illustrate in Figure 1, the accuracy of such an approximation heavily depends on the choice of the constant ω.



a) $\tau_N = 50$

b) $\tau_N = 500$

Figure 1 Performance of the approximation for the server utilization based on the limit result from (4) for various values of ω when $\mu_o = 250$, and $\lambda_o = 248$.

The intuition behind the crucial role of ω becomes more clear when we consider the extreme scenario where the firm hides infinite number of waiting spots, which corresponds to an infinitely large κ. Under this extreme scenario, limit result from (4) would become just ω. Therefore, we would propose $1 - \omega/\mu_o$ as the approximation for the server utilization if we wanted to use the results from Proposition 1. It is also important to note that the system would behave as a single-server queue with unlimited waiting room in this extreme scenario of infinitely large κ. Based on these two observations, one can notice that $1 - \omega/\mu_o$ should be equal to the fraction of time that an $M/M/1$ queue remains busy, which is λ_o/μ_o, to obtain an accurate approximation. Consequently, we pick $\mu_o - \lambda_o$ as the constant ω and use the following approximations for the server utilization and the average queue length in the firm's service facility:

$$\sigma(K,\lambda_o,\mu_o) \simeq \tilde{\sigma}(K,\lambda_o,\mu_o) \equiv 1 - \frac{\tilde{\omega}}{\mu_o}\left(1 + \frac{1}{e^{(K+\tau_N)\tilde{\omega}/\mu_o} - 1}\right) \quad (6)$$

$$Q(K,\lambda_o,\mu_o) \simeq \tilde{Q}(K,\lambda_o,\mu_o) \equiv \frac{\mu_o}{\tilde{\omega}} - \frac{K+\tau_N}{e^{(K+\tau_N)\tilde{\omega}/\mu_o} - 1} \quad (7)$$

where $\tilde{\omega} = \mu_o - \lambda_o$. As we demonstrate in Figure 2, the functions proposed above trace the server utilization and the average queue length functions from the original system with great accuracy for a wide range of K values. Hence, using the approximations in (6) and (7), we refine the firm's profit in the first period as

$$\Pi_1(K,\lambda_o,\mu_o) \simeq \tilde{\Pi}_1(K,\lambda_o,\mu_o) \equiv V * \mu * \tilde{\sigma}(K,\lambda_o,\mu_o) + h * \tilde{Q}(K,\lambda_o,\mu_o).$$

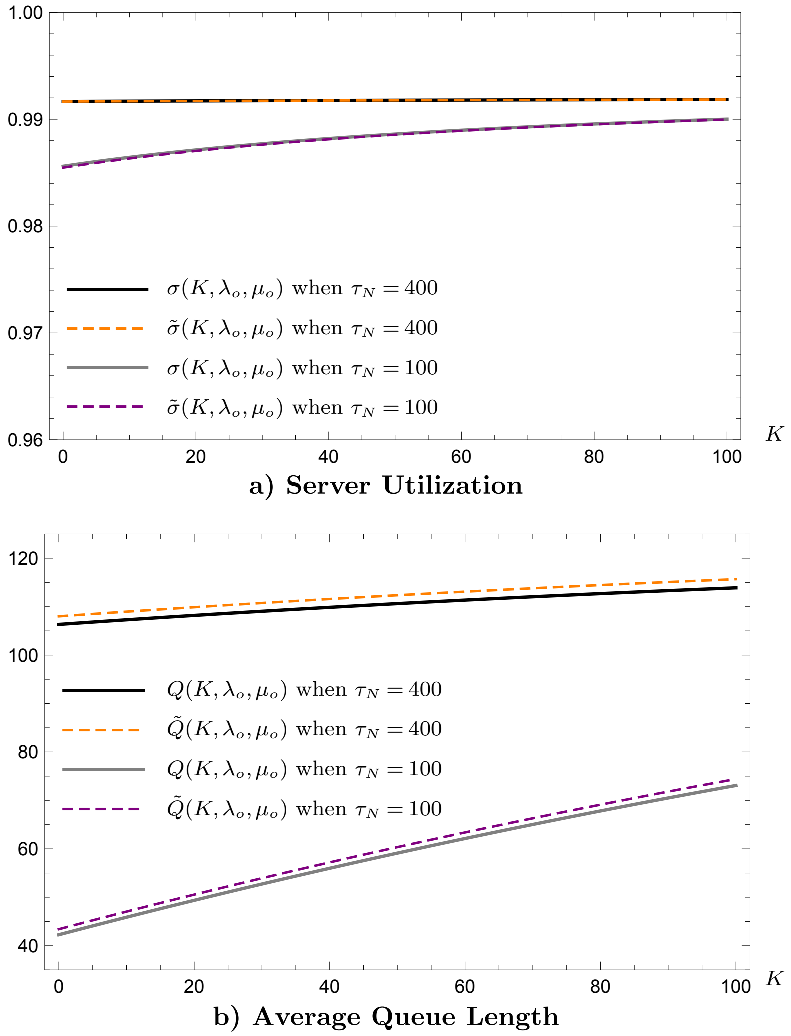
4.2. The Second Period

Once we establish our approximations for the first period, we now turn our attention to the firm's profits in the second period. As we mentioned in Section 3, the effective demand in the second period will be smaller than the potential demand because some of the strategic customers from the potential demand pool will not revisit the firm in the second period to penalize the disguised queues. Thus, it is crucial to first derive the effective arrival rate in the second period. Similar to our analysis for the server utilization and average queue length, we construct an approximation for the effective arrival rate in the second period by studying the sequence of the replica systems introduced at the beginning of Section 4. As a first step, we show how the function representing the fraction of customers not revisiting the firm, which is $\gamma\beta(K,\lambda,\mu)$, behaves along this sequence as the system size grows. We formally present our findings in Proposition 2.

Proposition 2. *Consider the sequence of systems introduced at the beginning of Section 4. If $\mu_n \to \infty$ as $n \to \infty$, then we have that*

$$\lim_{n\to\infty} \gamma\beta(\kappa\mu_n,\lambda_n,\mu_n) = \gamma\frac{e^{\kappa\omega} - 1}{e^{(\kappa+\tau_N/\mu_o)\omega} - 1}. \quad (8)$$

Similar to our results in Proposition 1, the above proposition yields an accurate approximation for the effective demand in the second period when we set ω equal to $\lambda_o - \mu_o$. More importantly,



a) Server Utilization

b) Average Queue Length

Figure 2 **Accuracy of the approximations for the server utilization and the average queue length in the First Period when** $\mu_o=250$ **and** $\lambda_o=248$.

Proposition 2 establishes that the fraction of customers not revisiting the firm in the second period will be strictly positive if the firm hides its waiting lines. As a result, the firm's facility will be underloaded in the second period even if the potential arrival rate is very close to the service rate. When the service center is underloaded, the customers service decision in the second period becomes immaterial because the queue length never exceeds the Naor threshold τ_N due to high service rate. Then, all customers revisiting the firm in the second period request service. Consequently, in the second period, the firm's service center operates as an underloaded single-server system with unlimited waiting capacity. This is a critical observation for obtaining efficient approximations for the second period because the performance metrics of an underloaded single-server system have very simple structures. For instance, the server utilization and the average queue length are both simple functions of the system load, conventionally denoted by ρ. Namely, the server utilization is ρ, and the average queue length is $\rho^2/(1-\rho)$. Based on these simple functional forms, we construct the following approximations for the server utilization and the average queue length in the second period:

$$\sigma(K,\lambda_o[1-\gamma\beta(K,\lambda_o,\mu_o)],\mu_o) \simeq \tilde{\sigma}_2(K,\lambda_o,\mu_o) \equiv \rho_o\left(1-\gamma\frac{e^{K\tilde{\omega}/\mu_o}-1}{e^{(K+\tau_N)\tilde{\omega}/\mu_o}-1}\right)$$

$$Q(K,\lambda_o[1-\gamma\beta(K,\lambda_o,\mu_o)],\mu_o) \simeq \tilde{Q}_2(K,\lambda_o,\mu_o) \equiv \frac{\tilde{\sigma}_2(K,\lambda_o,\mu_o)^2}{1-\tilde{\sigma}_2(K,\lambda_o,\mu_o)},$$

where we again have that $\tilde{\omega}=\lambda_o-\mu_o$. As Figure 3 illustrates, the functions proposed above accurately approximate the server utilization and the average queue length functions in the second period of the original system for a wide range of disguised-queue size, K, values. In this section, we use these approximations to refine the firm's profit in the second period as follows:

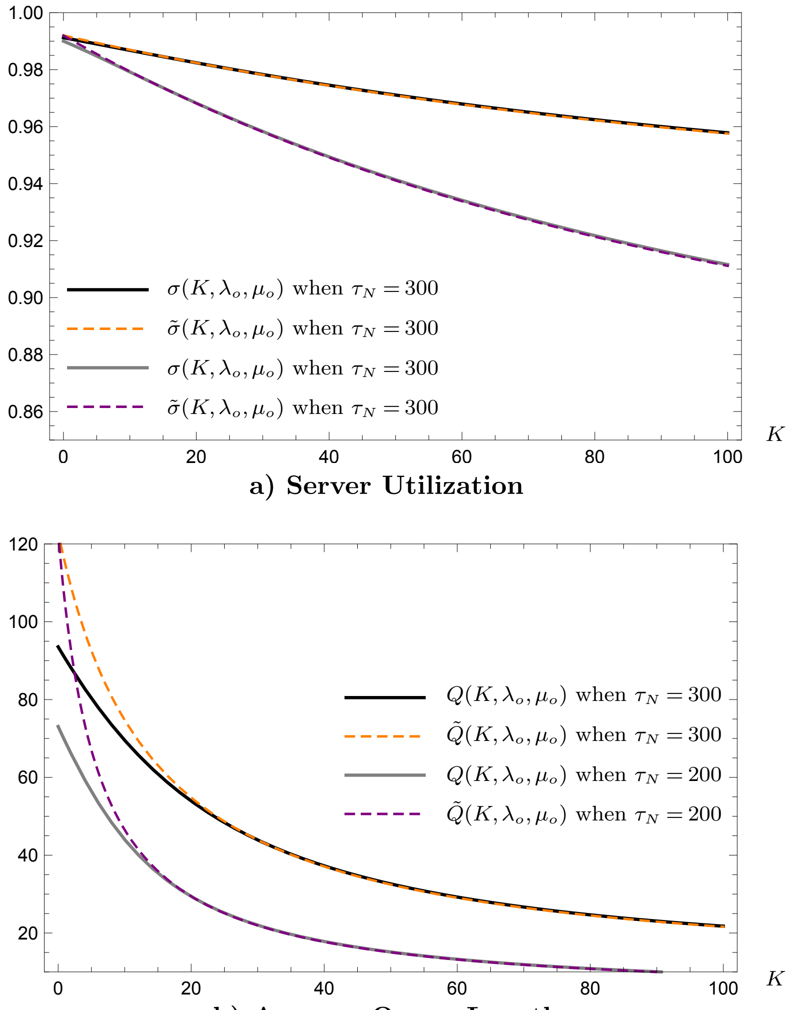
$$\Pi_2(K,\lambda_o,\mu_o) \simeq \tilde{\Pi}_2(K,\lambda_o,\mu_o) \equiv V*\mu*\tilde{\sigma}_2(K,\lambda,\mu)+h*\tilde{Q}_2(K,\lambda,\mu).$$

In this section, we build simple yet precise approximations for the key performance metrics of the firm's service facility. Then, we use these simplified forms of the metrics to refine the profit functions of the firm. Next, we solve the firm's queue disguising problem based on these refined functions.

5. The Queue Disguising Problem

Here, we study the firm's problem with the objective of finding optimal disguised-queue size. We carry out all of analysis based on the refined profit functions introduced in Section 4. We then numerically show that the analysis based on the approximated profit functions results in near optimal solutions for the firm's original problem.

As mentioned in the Introduction, the main goal of this paper is to study the implications of strategic behavior on the firm's queue disguising decision. To this end, we study the firm's problem in two different scenarios: i) none of the customers acts strategically ii) there is a strictly positive



a) Server Utilization

b) Average Queue Length

Figure 3 **Accuracy of the approximations for the server utilization and the average queue length in the Second Period when $\mu_o = 250$, $\lambda_o = 248$, and $\gamma = 2/3$.**

portion of customers who can strategically react to firm's queue disguising decision. We first focus on the scenario where customers are not strategic.

5.1. Optimal Decision with Naive Customers

As a first step towards understanding the role of strategic customer behavior on the firm's queue disguising strategy, we derive the firm's optimal decision when none of the customers are strategic. In other words, we start our analysis by setting the fraction of customers who can react to firm's decisions, which is denoted by γ, equal to zero. When customers are not strategic, the firm's profits functions in both periods become identical because the queue disguising decision has no ramifications on the second period demand. Hence, it is sufficient for the firm to consider only the first period profits while choosing its queue disguising strategy and solve the following problem, which we refer to as the *Naive Customers* problem:

$$\tilde{\Pi}^*_{nc} \equiv \max_{K\geq 0} \tilde{\Pi}_1(K, \lambda_o, \mu_o).$$

The firm's profit in the above problem is not always a concave function of the disguised-queue size, K. In fact, it can start as a concave function and then become convex. However, we show that the firm's (refined) first period profit is unimodal in the disguised-queue size, and thus the firm's problem when facing non-strategic customers has a unique maximizer. We formally present the optimal solution of the *Naive Customers* problem in the following theorem.

THEOREM 1. *The optimal disguised-queue size in the Naive Customers problem, denoted by K^*_{nc}, is*

$$K^*_{nc} = \max\{0, \mu_o x_{nc}/\tilde{\omega} - \tau_N\},$$

where x_{nc} is the unique non-negative solution for $e^x\left(1 + \frac{V\tilde{\omega}^2}{h\mu} - x\right) = 1$. K^*_{nc} is decreasing in the holding cost h and the Naor threshold, τ_N. Furthermore, there exist two non-negative critical values h_{nc} and $\tau_{N nc}$ such that

$$K^*_{nc} = 0 \textit{ for any } h \geq h_{nc} \textit{ or } \tau_N \geq \tau_{N nc}.$$

The above theorem establishes that the firm may find it optimal to hide a portion of its waiting line. When the firm starts to disguise some of its waiting spots, it makes customers to keep requesting service even when there are more people in the queue than the Naor threshold τ_N. As a result, the server utilization goes up because less customers leave the system without joining the queue. On the other hand, the firm also has to handle longer queues when it employs the queue disguising strategy. Evidently, the gains from improved server utilization outweigh the losses from having longer queues when the holding cost h and the Naor threshold τ_N are both low. Hence, the firm optimally chooses a positive disguised-queue size K. However, we also show that queue disguising strategy becomes less attractive as h and τ_N increases. In fact, once the holding cost or the Naor threshold exceeds a critical level, the firm optimally choose not to hide any parts of its waiting line.

Theorem 1 confirms the optimality of queue disguising strategy for low levels of holding cost and Naor threshold but it is still an open question whether the queue disguising strategy results in significant profit improvements. To answer that, we next asses how much the firm benefits when the optimal disguised-queue size is non-zero. Specifically, we define the relative profit gain from the queue disguising strategy as

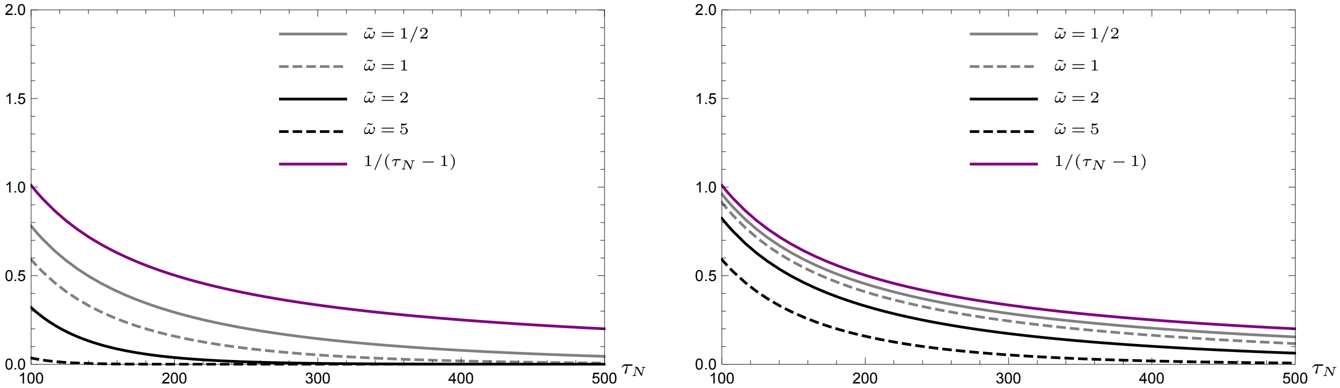
$$\mathcal{G}_{nc} = 100 \times \left(\frac{\tilde{\Pi}^*_{nc}}{\tilde{\Pi}_1(0, \lambda_o, \mu_o)} - 1 \right). \tag{9}$$

Proposition 3. *Let $\mathcal{G}_{nc}$ be the relative profit gain from the queue disguising strategy as defined in (9). Then, the following statements holds true:*

1. *$\mathcal{G}_{nc}$ is decreasing in τ_N.*
2. *if $\tau_N \geq 4$ and $\mu_o \geq 2\tilde{\omega}$, then $\mathcal{G}_{nc}$ is decreasing in h. Furthermore, we have that*

$$\mathcal{G}_{nc} \leq \overline{\mathcal{G}}_{nc} \equiv \lim_{h \to 0} \mathcal{G}_{nc} \leq \frac{1}{\tau_N - 1}.$$

Aligned with the findings in Theorem 1, the above proposition shows that the benefits from the queue disguising strategy decline as the Naor threshold or the holding cost increases. This is somewhat expected because the optimal disguised-queue size also decreases in these two parameters. More interestingly, we show that the profits gains from the queue disguising strategy is bounded from above. Although this upper bound has a complex structure from which it is hard to derive direct insights, we show that the bound is always smaller than a simple function of the Naor threshold, $1/(\tau_N - 1)$ to be specific. An important implication of this relationship is that the firm's optimal queue disguising strategy can only lead to negligible profit improvements in service system where it is normal for customers to wait hundreds of people before their services commence. As illustrated in Figure 4, the upper bound we derive for the profit gains from queue disguising, $\overline{\mathcal{G}}_{nc}$, can be as small as 1% in systems with high levels of Naor thresholds. Figure 4 also highlights that the bound $\overline{\mathcal{G}}_{nc}$ is close to $1/(\tau_N - 1)$ only when the system load of the firm's service facility is nearly 1, which occurs when $\tilde{\omega}$ is small relative to the service rate μ_o. $\overline{\mathcal{G}}_{nc}$ can be significantly smaller than $1/(\tau_N - 1)$ if the system load is away from 1.



a) $\mu_o = 100$

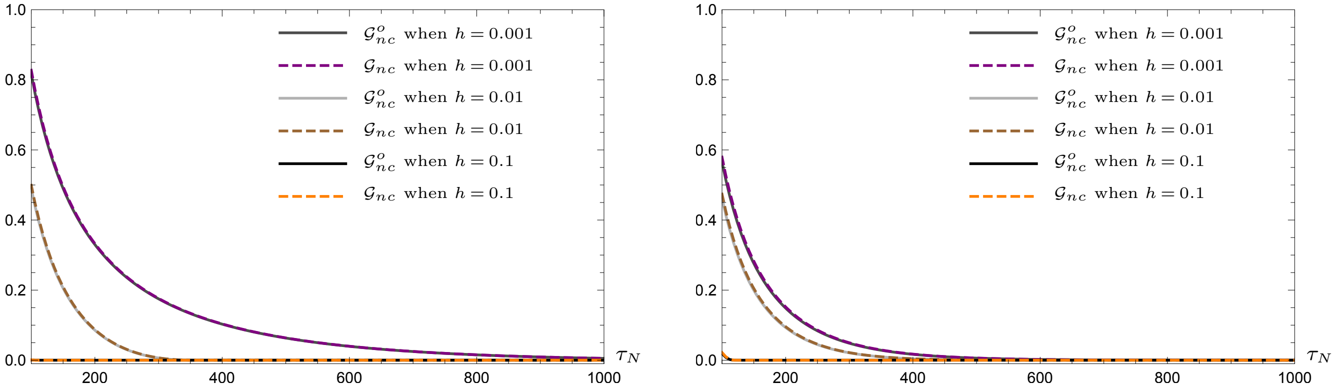
b) $\mu_o = 500$

Figure 4 The upper bound of the relative profit gain from the queue disguising strategy, $\overline{\mathcal{G}}_{nc}$, when $V = 1$.

We conclude our analysis of the firm's queue disguising problem in the absence of strategic customers by demonstrating the accuracy of the optimal disguised-queue size we derive based on the refined profit function. As a first step towards this goal, we numerically solve the firm's original problem for various parameters and find the exact optimal disguised-queue size, denoted by K^o_{nc}. We then calculate the relative profit gains from queue disguising strategy in the firm's original problem as

$$\mathcal{G}^o_{nc} \equiv 100 \times \left(\frac{\Pi_1(K^o_{nc}, \lambda_o, \mu_o)}{\Pi_1(0, \lambda_o, \mu_o)} - 1 \right).$$

We present the results of our numerical study in Figure 5.



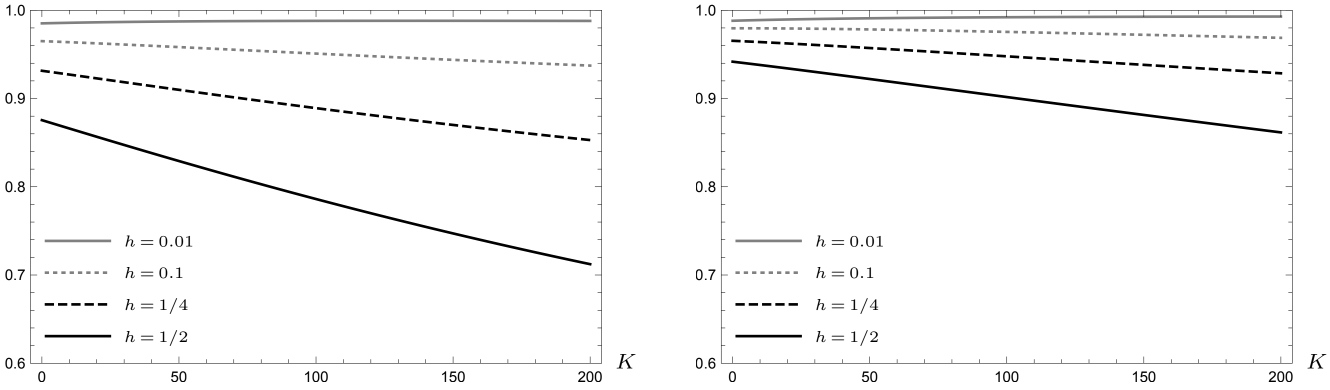
a) $\tilde{\omega} = 1$

b) $\tilde{\omega} = 5$

Figure 5 The relative profit gain from the queue disguising strategy, $\mathcal{G}_{nc}$, when $\mu_o = 500$ and $V = 1$.

It is clear from Figure 5 that the analysis we carry out using the refined profit functions generates exceptionally precise approximation for the benefits the firm obtain from the queue disguising strategy. The above figure also shows that the upper bound $1/(\tau_N - 1)$ is achieved only for extremely low values of the holding cost h. In fact, the upper bound is achieved only when h is fractional (e.g., less than 0.1%) compared to V, the value that firm earns from serving customers.

Our numerical study also reveals that misuse of the queue disguising strategy can be costly for a firm facing non-strategic customers, especially when disguised queues are not preferable. We observe that the firm's profit may decline sharply as the disguised-queue size K increases if queue disguising is not optimal. Interestingly, the firm's profit, as a function of K, seems flat for the cases where the firm optimally hides its queues. The main driver of this result is that the marginal changes in the firm's profit as K changes highly depend on the holding cost h. Thus, the optimality of queue disguising, which occurs when holding cost is trivially small, implies that the firm's profit is actually almost unresponsive to the changes in the disguised-queue size. On the contrary, the fact that queue disguising is undesirable, which happens under moderate to high holding costs, indicates a steep decline in the firm's profit as the disguised-queue size increases. We illustrate that observation in Figure 6.



a) $\mu_o = 200$ **b)** $\mu_o = 500$

Figure 6 **The firm's profit in the absence of strategic customer as a function of the disguised-queue size K when $\tilde{\omega} = 1$ and $V = 1$. The profit is divided by μ_o to have the same scale in both figures.**

After reviewing the *Naive Customers* problem, we turn our attention the queue disguising decision of a firm facing strategic customers.

5.2. Optimal Decision with Strategic Customers

We continue our analysis by studying the firm's problem in the presence of strategic customers. As the firm faces strategic customers, the firm's profits in both periods are no longer identical. Hence, the firm has to consider both the first and the second period profit functions and solve the following problem:

$$\tilde{\Pi}^*_{sc} \equiv \max_{K \geq 0} \tilde{\Pi}_1(K, \lambda_o, \mu_o) + \tilde{\Pi}_2(K, \lambda_o, \mu_o).$$

We refer to the above problem as the *Strategic Customers* problem. Unlike, the *Naive Customers* problem, the above problem is not amenable to use standard first-order conditions in order to find an optimal solution. Therefore, rather than deriving the exact solution of the *Strategic Customers* problem, we attempt to obtain the structural properties of the firm's optimal queue disguising decision in systems where the original service rate μ_o is large. We formally present our findings in the following theorem.

Theorem 2. *The optimal disguised-queue size in the Strategic Customers problem, denoted by K^*_{sc}, satisfies that*

$$\lim_{\mu_o \to \infty} \frac{K^*_{sc}}{\hat{K}_{sc}} = 1,$$

where

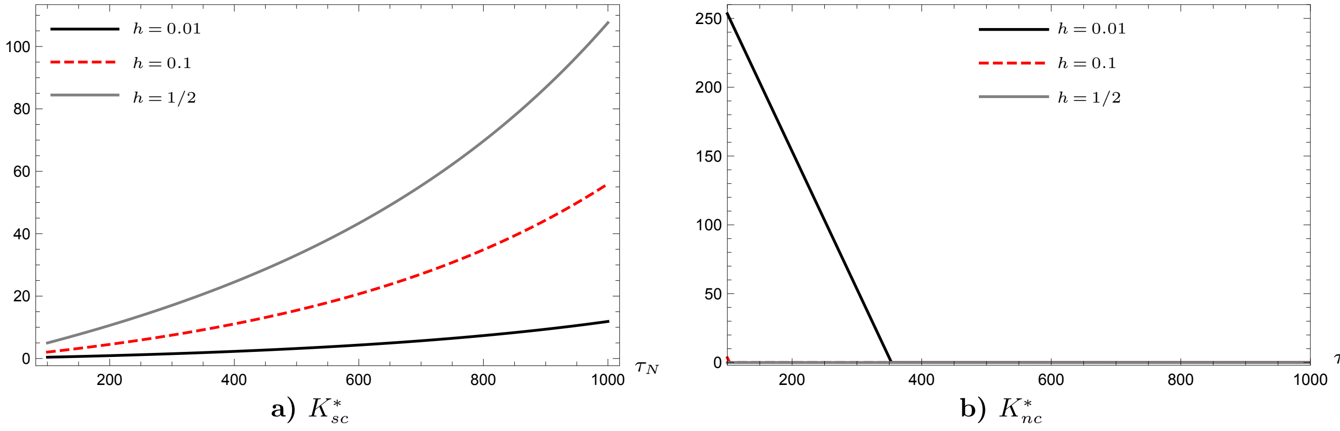
$$\hat{K}_{sc} = \frac{\left(e^{\tau_N \tilde{\omega}/\mu_o} - 1\right)\left(\sqrt{hV\mu_o} - V\tilde{\omega}\right)}{\gamma V \tilde{\omega}}.$$

Furthermore, $\hat{K}_{sc}$ is increasing in the holding cost h and the Naor threshold.

The above theorem shows that the optimal disguised-queue size is asymptotically equivalent to a relatively simple expression, $\hat{K}_{sc}$. As an implication of this asymptotic relationship, we expect that the firm's optimal disguised-queue size in the *Strategic Customers* problem, denoted by K^*_{sc}, will exhibit similar structural properties to $\hat{K}_{sc}$ if the service speed is sufficiently high. Theorem 2 shows that $\hat{K}_{sc}$ is a monotone increasing function of the holding cost and the Naor threshold, thereby implying that the optimal disguised-queue size should also be an increasing function of h and τ_N in sufficiently fast systems.

Our findings in Theorem 2 suggest a significant shift in how the firm utilizes the disguised queues.

Specifically, the above results suggest that the firm has to hide more waiting spots as the holding cost or the Naor threshold increases while facing strategic customers. However, this completely contrast with the firm's optimal decisions in the *Naive Customers* problem, where the optimal disguised-queue size declines in h or τ_N. We illustrate this contrast in Figure 7, where we obtain K^*_{sc} by numerically solving the *Strategic Customers* problem.



a) K^*_{sc} b) K^*_{nc}

Figure 7 The comparison of the firm's optimal queue disguising decisions in the Strategic Customers and Naive Customers problems when $V = 1$, $\mu_o = 500$, $\tilde{\omega} = 1$, and $\gamma = 2/3$.

The main driver of the U-turn in firm's queue disguising strategy due to strategic customers is that in the *Strategic Customers* problem, the future customer arrivals depend on the firm's queue disguising decision. In contrast, in the absence of the strategic behavior, the firm's queue disguising decision in the first period has no impact on the future demand. Hence, the disguised queues become undesirable in the *Naive Customers* problem once the cost of longer queues they cause outweighs the benefits of higher server utilization achieved owing to deceiving customers. This occurs when holding cost or Naor threshold is high. However, as two periods are interconnected in the *Strategic Customers* problem, the firm's queue disguising decision will play an important role to moderate the effective demand in the second period besides just affecting customers' service decisions in the first period. To be specific, by hiding its waiting lines, the firm can curb some of the future demand (if needed) and reduce the customer waiting and average queue length in the second period. It turns out, when holding cost or the Naor threshold is high, running a second period with shorter queues at the expense of longer queues in the first period is more appealing for the firm than having two moderately loaded systems. As the comparison between Figures 2.b and 3.b illustrates, the average queue length in the first period increases in the disguised-queue size K in a linear fashion whereas the queues in the second period declines by K at an exponential rate. This difference in how the average queue length in each period responds to the changes in disguised-queue size is one of the main rationale behind a non-zero optimal disguised-queue size.

Similar to the impact of strategic customer behavior on the optimal disguised-queue size, we find that the structure of the profit gains from disguised queues also alters as the customers becomes strategic. Specifically, we show that the firm obtains more benefits from disguised queues as the holding cost increases in the *Strategic Customers* problem when the service rate is sufficiently high. We also show that the benefits from queue disguising strategy increases by the Naor threshold τ_N as long as τ_N is not too high compared to the service rate. Although the firm's revenue improvements due to queue disguising may decline as the Naor threshold rises, it is important to note that the disguised queues always benefit the firm. Recall that, in Proposition 3, we show that the benefits from the optimal queue disguising strategy always decreases by h and τ_N. Hence, the strategic reaction of customers also has a significant impact on the relationship between the firm's benefits from disguised queues and the holding cost (or the customer patience). We present these results formally in Proposition 4 by defining the relative profit gain from the optimal queue disguising strategy in the *Strategic Customers* problem as

$$\mathcal{G}_{sc} = 100 \times \left(\frac{\tilde{\Pi}^*_{sc}}{2\tilde{\Pi}_1(0, \lambda_o, \mu_o)} - 1 \right). \tag{10}$$

Proposition 4. *Let $\mathcal{G}_{sc}$ be the relative profit gain from the queue disguising strategy as defined in (10). Letting $\zeta \equiv \lim_{\mu_o \to \infty} \tau_N/\mu_o$, we have that*

$$\lim_{\mu_o \to \infty} \mathcal{G}_{sc} = \tilde{\mathcal{G}}_{sc}(h, \zeta),$$

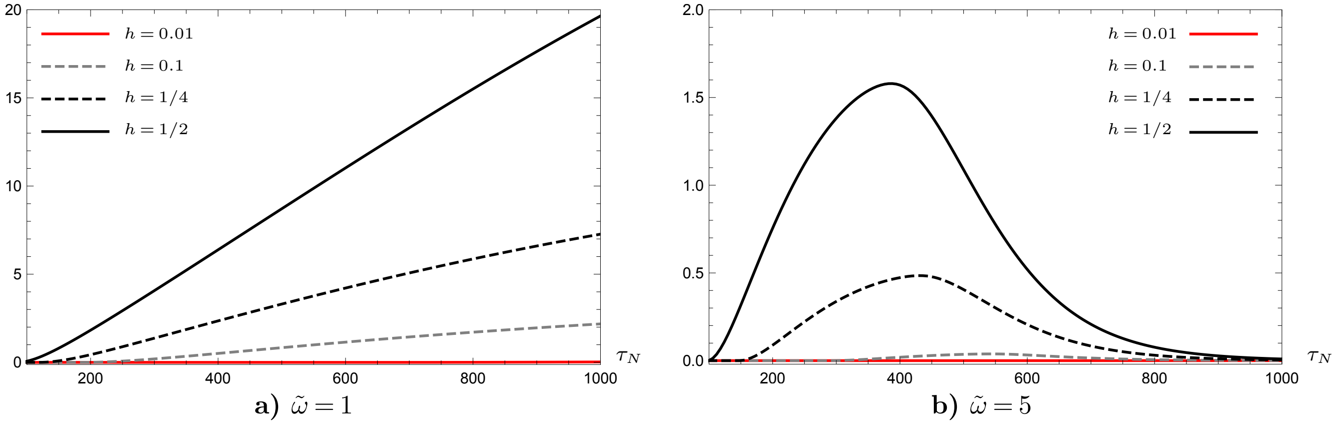
where $\tilde{\mathcal{G}}_{sc}(h, \zeta)$ is an increasing function of h and ζ.

Furthermore, for any fixed μ_o, we have that

$$\lim_{\tau_N \to \infty} \mathcal{G}_{sc} = \frac{h\tilde{\omega}(2\mu_o - \tilde{\omega})}{2\mu_o(V\tilde{\omega}(\mu_o - \tilde{\omega}) - h\mu_o)} > 0.$$

Once we obtain the above analytical results based on the refined forms of the profit functions, we again turn our attention to the firm's original problem. As in the case of non-strategic customers, we numerically solve the firm's original problem for various parameters and find the exact optimal disguised-queue size, denoted by K^o_{sc}. We then and calculate the relative profit gains from queue disguising strategy in the firm's original problem as

$$\mathcal{G}^o_{sc} \equiv 100 \times \left(\frac{\Pi_1(K^o_{sc}, \lambda_o, \mu_o) + \Pi_2(K^o_{sc}, \lambda_o, \mu_o)}{2\Pi_1(0, \lambda_o, \mu_o)} - 1 \right).$$



a) $\tilde{\omega} = 1$ **b)** $\tilde{\omega} = 5$

Figure 8 **The relative profit gain from the queue disguising strategy, $\mathcal{G}^o_{nc}$, when $\mu_o = 500$, $V = 1$, and $\gamma = 2/3$.**

The above figures, which present the results of our numerical study, verifies that the insights we obtain based on our asymptotic analysis also hold in the firm's original problem. Figure 8 also illustrates how the strategic behavior of customers alters the effectiveness of queue disguising when we compare it with Figure 5. The comparison of these figures shows that for low levels of holding cost, the firm's benefits from queue disguising decline as the customers start to react to disguised queues. More specifically, the relative profit gains $\mathcal{G}^o_{sc}$ is virtually zero in the above figure when $h = 0.01$ whereas the gains under non-strategic customers $\mathcal{G}^o_{nc}$ is about 1%, a small yet non-zero, in Figure 5. More strikingly, when the holding costs are high (e.g. more than 10% of the firm's reward V), the firm starts to enjoy considerable benefits from queue disguising strategy although the queue disguising yields almost zero gains in the absence of strategic customers. Given that the service environments we consider in this paper incur non-negligible holding costs, our analysis advises these services to take advantage of disguised queues only if customers are bothered by making erroneous decisions due to disguised waiting lines. When customers are non-strategic or do not react to queue disguising, we suggest firms to make their entire waiting line visible because any disguised queue will be costly for them.

6. Conclusion

Service systems commonly suffer from the mismatch between demand and supply in the forms of server idleness and the waiting lines. Hence, service providers are often in search for alleviating one of these issues without exacerbating the other one. As the waiting line management may require handling of customers' decisions, it is naturally the harder problem, and thus received more attention both in academia and the "real world". In this paper, we contribute to the growing literature on waiting line management. In particular, we study a firm that can partially hide the waiting line forming in front of its service center, thereby pushing customers to make their decisions based on the observable portion of the queue. We also suppose that some of the customers are strategic and penalize the firm by not revisiting the firm in the future if they are fooled to request service because of the disguised queues.

The main focus of this paper is to gain insights about how the strategic behavior of the customers impact the firm's queue disguising decisions. Hence, we first focus on the case where customers are not strategic. We show that the queue disguising is optimal for a firm only when the holding cost and the customers' tolerance for waiting are low. Otherwise, the firm does not find it optimal to hide any of its waiting spots. In the absence of strategic customers, we also find that the firm's optimal queue disguising strategy yields only negligible profit improvements, relative to letting customers observe the entire waiting line. In fact, these minimal profit improvements vanish as holding customers in the queue becomes costlier. On the other hand, when some of the customers behave strategically, we find that the optimally chosen level of disguised queue can increase the firm's profit considerably. More interestingly, the firm's profit gains from disguised queues increase as the holding cost increases unlike the case without strategic customers. This result brings to light a crucial insight for the service environments we consider in this paper as these systems bear non-negligible costs to keep the waiting customers happy: The firms can significantly benefit from disguised queues if customers act strategically and penalize the firm by not revisiting the firm whereas any disguised queues hurt a firm serving non-strategic customers.

As the service facilities we review in this paper usually attract a high volume of demand and serve customers very rapidly, we solve the firm's queue disguising problem using an approximation built on the asymptotic behavior of the firm's original service center. Specifically, we consider a sequence of systems that are the replicas of the firm's service center in a parametric regime where the demand and the arrival rates grow unboundedly. We illustrate that our asymptotic analysis lead to efficient and accurate approximations for many crucial system metrics such as server utilization and average queue length. We, then, use these approximations to refine the firm's profit functions. Through an extensive numerical study, we support the robustness of our findings that are obtained based on the refined profit function.

References

Afeche, F., H. Mendelson. 2004. Pricing and priority auctions in queueing systems with generalized delay cost structure. *Management Sci.* **50** 869–882.

Allon, G., A. Bassamboo, I. Gurvich. 2011. “we will be right with you”: Managing customer expectations with vague promises and cheap talk. *Operations research* **59**(6) 1382–1394.

Allon, G., A. Federgruen. 2008. Service competition with general queueing facilities. *Operations Research* **56** 827–849.

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

Boudali, O., A. Economou. 2012. Optimal and equilibrium balking strategies in the single server markovian queue with catastrophes. *European Journal of Operational Research* **218**(3) 708–715.

Cachon, G., F. Zhang. 2007. Obtaining fast service in a queuing system via performance-based allocation of demand. *Management Science* **53**(3) 408–420.

Chiu, Chun-Hung, Tsan-Ming Choi, Yongjian Li, Lei Xu. 2014. Service competition and service war: a game-theoretic analysis. *Service Science* **6**(1) 63–76.

De Vany, A. 1976. Uncertainty, waiting time, and capacity utilization: A stochastic theory of product quality. *Journal of Political Economy* **84** 523–540.

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

Economou, A., S. Kanta. 2008. Optimal balking strategies and pricing for the single server markovian queue with compartmented waiting space. *Queueing Systems* **59**(3-4) 237–269.

Fitzsimmons, J., M. Fitzsimmons, S. Bordoloi. 2014. *Service management: Operations, strategy, information technology*. McGraw-Hill Higher Education.

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

Hassin, R. 1986. Consumer information in markets with random product quality: The case of queues and balking. *Econometrica: Journal of the Econometric Society* 1185–1195.

Hassin, R., M. Haviv. 2003. *To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems*. Kluwer Academic Publishers, Boston, MA.

Jouini, O., Z. Aksin, Y. Dallery. 2011. Call centers with delay information: Models and insights. *Manufacturing & Service Operations Management* **13**(4) 534–548.

Kim, B., J. Kim. 2017. Optimal information disclosure policies in a strategic queueing model. *Operations Research Letters* **45**(2) 181–186.

Li, L., L. Jiang, L. Liu. 2012. Service and price competition when customers are naive. *Production and Operations Management* **21**(4) 747–760.

Mendelson, H., S. Whang. 1990. Optimal incentive-compatitable priority pricing for the M/M/1 queue. *Operations Research* **38** 870–883.

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

Rubin, J. 2016. TEA/AECOM 2015 Theme Index and Museum Index: The Global Attractions Attendance Report. Attendance report, Themed Entertainment Association (TEA). URL `http://www.teaconnect.org/images/files/TEA_160_611852_160525.pdf`.

Simhon, Eran, Yezekael Hayel, David Starobinski, Quanyan Zhu. 2016. Optimal information disclosure policies in strategic queueing games. *Operations Research Letters* **44**(1) 109–113.

Whitt, Ward. 1999. Improving service by informing customers about anticipated delays. *Management science* **45**(2) 192–207.

Ziani, S., F. Rahmoune, M. S. Radjef. 2015. Customers'strategic behavior in batch arrivals $\mathrm{M}^2/\mathrm{M}/1$ queue. *European Journal of Operational Research* **247**(3) 895–903.

Appendix A: Proofs in Section 4

A.1. Proof of Proposition 1

We first want to note that the server utilization function, $\sigma(K,\lambda,\mu)$, in an $M/M/1/K+R\mu/c$ queue with arrival rate λ and service rate μ can be written as

$$\sigma(K,\lambda,\mu)=1-\frac{1-\rho}{1-\rho^{K+R\mu/c+1}}$$

where $\rho\equiv\lambda/\mu$.

Using the above equation, we have that

$$\big[1-\sigma(\kappa\mu_n,\lambda_n,\mu_n)\big]\mu_n=\mu_n\frac{1-\rho_n}{1-\rho_n^{(\kappa+\tau_N/\mu_o)\mu_n+1}}=\frac{\omega}{1-(1-\omega/\mu_n)^{(\kappa+R/c)\mu_n+1}}.$$

Then, our claim in the proposition holds by using the fact that

$$\lim_{n\to\infty}(1-\omega/\mu_n)^{(\kappa+R/c)\mu_n+1}=e^{-(\kappa+R/c)\omega}.$$

Regarding our claim about the average queue length, the server utilization function, $Q(K,\lambda,\mu)$, in an $M/M/1/K+R\mu/c$ queue with arrival rate λ and service rate μ can be written as

$$Q(K,\lambda,\mu)=K+\frac{R\mu}{c}+\frac{\rho}{1-\rho}-\frac{K+R\mu/c+\rho}{1-\rho^{K+R\mu/c+1}}$$

Using the above equation, we have that

$$\begin{aligned}\big[Q(\kappa\mu_n,\lambda_n,\mu_n)\big]/\mu_n &= (\kappa+R/c)\frac{\rho^{K+R\mu_n/c+1}}{1-\rho^{K+R\mu_n/c+1}}+\frac{\rho_n}{\mu_n(1-\rho_n)}-\frac{\rho_n}{\mu_n[1-\rho^{K+R\mu/c+1}]}\\ &= (\kappa+R/c)\frac{\rho^{K+R\mu_n/c+1}}{1-\rho^{K+R\mu_n/c+1}}+\frac{1}{\omega}-\frac{\rho_n}{\mu_n[1-\rho^{K+R\mu/c+1}]}\end{aligned}$$

Then, our claim in the proposition holds by using the fact that

$$\lim_{n\to\infty}\rho_n^{(\kappa+R/c)\mu_n+1}=e^{-(\kappa+R/c)\omega}.$$

A.2. Proof of Proposition 2

The fraction of customers not revisiting, $\gamma\beta(K,\lambda,\mu)$, in an $M/M/1/K+R\mu/c$ queue with arrival rate λ and service rate μ can be written as

$$\gamma\beta(K,\lambda,\mu)=\gamma\rho^{R\mu/c}\frac{1-\rho^K}{1-\rho^{K+R\mu/c+1}} \tag{11}$$

where $\rho\equiv\lambda/\mu$. Using the above equation, we have that

$$\lim_{n\to\infty}\gamma\beta(\kappa\mu_n,\lambda_n,\mu_n)=\gamma\frac{e^{-R\omega/c}-e^{-(\kappa+R/c)\omega}}{1-e^{-(\kappa+R/c)\omega}}=\gamma\frac{e^{\kappa\omega}-1}{e^{(\kappa+R/c)\omega}-1},$$

where the first equality holds since

$$\lim_{n\to\infty}\rho_n^{x\mu_n}=e^{-x\omega},$$

for any constant x.

Appendix B: Proofs in Section 5

B.1. Proof of Theorem 1

We first want to note that the firm's refined profit function in the first period is

$$\tilde{\Pi}_1(K,\lambda_o,\mu_o) = \mu_o\left(V - \frac{h}{\tilde{\omega}}\right) - V\tilde{\omega} + \frac{chK - cV\tilde{\omega} + h\mu_o R}{c\left(e^{\frac{R\tilde{\omega}}{c}+\frac{K\tilde{\omega}}{\mu_o}} - 1\right)}. \tag{12}$$

By taking the derivative of $\tilde{\Pi}_1(K,\lambda_o,\mu_o)$ with respect to K, we have that

$$\begin{aligned}\frac{d\tilde{\Pi}_1(K,\lambda_o,\mu_o)}{dK} &= \frac{e^{\frac{R\tilde{\omega}}{c}+\frac{K\tilde{\omega}}{\mu_o}}(ch\mu_o - h\tilde{\omega}(cK+\mu_o R) + cV\tilde{\omega}^2) - ch\mu_o}{c\mu_o\left(e^{\frac{R\tilde{\omega}}{c}+\frac{K\tilde{\omega}}{\mu_o}} - 1\right)^2}\\ &= \frac{h\left(e^x(A-x)-1\right)}{\left(e^x-1\right)^2},\end{aligned}$$

where the second equality is obtained by defining $x \equiv \frac{R\tilde{\omega}}{c} + \frac{K\tilde{\omega}}{\mu_o}$ and $A \equiv 1 + \frac{V\tilde{\omega}^2}{h\mu_o}$.

Using the above derivative, we, first, argue that the profit function $\tilde{\Pi}_1(K,\lambda_o,\mu_o)$ is unimodal. To this end, we let $f(x) \equiv e^x(A-x)-1$. Notice that $f(0) = A-1 > 0$. Furthermore, we have that $f'(x) = e^x(A-x-1)$, so that $f(x)$ is increasing for any $x < A-1$ and decreasing otherwise. Combining these observations with the fact that $lim_{x\to\infty} f(x) = -\infty$, we obtain that $f(x) = 0$ has a unique root, which is referred to as x_{nc}, and $f(x)$ is positive for any $x < x_{nc}$ and negative otherwise. This also proves that $\tilde{\Pi}_1(K,\lambda_o,\mu_o)$ is increasing for any $K < \mu_o x_{nc}/\tilde{\omega} - R\mu_o/c$ and decreasing otherwise since the derivative of $\tilde{\Pi}_1(K,\lambda_o,\mu_o)$ has the same sign as $f(x)$.

B.2. Proof of Proposition 3

We first want to note that we can rewrite the refined first period profit function, as given in (12), after substituting $\frac{R\tilde{\omega}}{c} + \frac{K\tilde{\omega}}{\mu_o}$ by x as follows:

$$\begin{aligned}\pi_1(x,\lambda_o,\mu_o) &= \frac{h\mu_o\left(x - e^x + 1\right) + V\tilde{\omega}\left[(e^x-1)\mu_o - \tilde{\omega}e^x\right]}{\tilde{\omega}\left(e^x-1\right)}\\ &= V\mu_o s(x,\mu_o/\tilde{\omega}) - hq(x,\mu_o/\tilde{\omega}),\end{aligned}$$

where $s(x,\theta) = 1 - \frac{e^x}{\theta(e^x-1)}$ and $q(x,\theta) = \theta - \frac{\theta x}{e^x-1}$.

We also want to note that for the monotonicity results of $\mathcal{G}_{nc}$, it is sufficient to focus only on the cases where $K^*_{nc} > 0$, i.e. $x_{nc} > x_0 \equiv \tau_N\tilde{\omega}/\mu_o$. Therefore, using our results from Theorem 1, we have that $\tilde{\Pi}^*_{nc} = \pi_1(x_{nc},\lambda_o,\mu_o)$. We also have that $\tilde{\Pi}_1(0,\lambda_o,\mu_o) = \pi_1(x_0,\lambda_o,\mu_o)$.

1. Note that the function $\pi_1(x,\lambda_o,\mu_o)$ does not depend on the Naor threshold τ_N. Furthermore, x_{nc} is also independent of τ_N. Therefore, any change in the parameter τ_N does not affect the optimal profit $\tilde{\Pi}^*_{nc}$. On the other hand, $\pi_1(x,\lambda_o,\mu_o)$ is an increasing function of x for any $x < x_{nc}$ by the optimality of x_{nc}. This implies that $\pi_1(x_0,\lambda_o,\mu_o)$ increases by τ_N since we focus on the cases where $x_{nc} > x_0$. Combining these two observations, we have that $\frac{\tilde{\Pi}^*_{nc}}{\tilde{\Pi}_1(0,\lambda_o,\mu_o)}$ is decreasing in τ_N, which directly implies that $\mathcal{G}_{nc}$ is also decreasing in τ_N.

2. By taking the derivative of $\mathcal{G}_{nc}$ with respect to h, we have that

$$\begin{aligned}\frac{d\mathcal{G}_{nc}}{dh} &= \frac{d\pi_1(x_{nc},\lambda_o,\mu_o)}{dh}\frac{1}{\pi_1(x_0,\lambda_o,\mu_o)} - \frac{d\pi_1(x_0,\lambda_o,\mu_o)}{dh}\frac{\pi_1(x_{nc},\lambda_o,\mu_o)}{\pi_1(x_0,\lambda_o,\mu_o)^2}\\ &= \frac{V\mu_o\tilde{\omega}s(x_{nc},\mu_o/\tilde{\omega})s(x_0,\mu_o/\tilde{\omega})}{\pi_1(x_0,\lambda_o,\mu_o)^2}\left[\frac{q(x_0,\mu_o/\tilde{\omega})}{s(x_0,\mu_o/\tilde{\omega})} - \frac{q(x_{nc},\mu_o/\tilde{\omega})}{s(x_{nc},\mu_o/\tilde{\omega})}\right],\end{aligned}$$

where $s(x,\theta)$ and $q(x,\theta)$ are as defined at the beginning of the proof.

Note that $\frac{q(x,\theta)}{s(x,\theta)} = \frac{\theta^2(-x+e^x-1)}{e^x(\theta-1)-\theta}$, and thus $\frac{q(x,\theta)}{s(x,\theta)}$ is increasing in x for any $x \geq \tau_N/\theta$, $\tau_N \geq 4$, and $\theta \geq 2$ by Lemma 2. Thus, defining $\theta \equiv \mu_o/\tilde{\omega}$, we have that

$$\frac{q(x_0,\mu_o/\tilde{\omega})}{s(x_0,\mu_o/\tilde{\omega})} - \frac{q(x_{nc},\mu_o/\tilde{\omega})}{s(x_{nc},\mu_o/\tilde{\omega})} < 0,$$

for any $\tau_N \geq 4$ and $\mu_o/\tilde{\omega} \geq 2$, which completes the proof for the monotonicity of $\mathcal{G}_{nc}$ in h.

As a direct implication of the the monotonicity of $\mathcal{G}_{nc}$ in h and the fact that $\lim\limits_{h\to 0} x_{nc} = \infty$, we have that

$$\mathcal{G}_{nc} \leq \overline{\mathcal{G}}_{nc} \equiv \lim_{h\to 0}\frac{\lim\limits_{x\to\infty}\pi_1(x,\lambda_o,\mu_o)}{\pi_1(x_0,\lambda_o,\mu_o)} - 1 = \frac{\tilde{\omega}}{e^{\frac{\tau_N\tilde{\omega}}{\mu_o}}(\mu_o - \tilde{\omega}) - \mu_o}.$$

By taking the derivative of $\overline{\mathcal{G}}_{nc}$ with respect to μ_o, we have that

$$\frac{\partial\overline{\mathcal{G}}_{nc}}{\partial\mu_o} = \frac{\tilde{\omega}\left(1 - e^{\frac{\tau_N\tilde{\omega}}{\mu_o}}(1 - \tau_N\tilde{\omega}/\mu_o + \tau_N\tilde{\omega}^2/\mu_o^2)\right)}{\left((\tilde{\omega}-\mu_o)e^{\frac{\tau_N\tilde{\omega}}{\mu_o}} + \mu_o\right)^2} = \frac{\tilde{\omega}\left(1 - e^{\frac{\tau_N}{\theta}}(1 - \tau_N/\theta + \tau_N/\theta^2)\right)}{\left((\tilde{\omega}-\mu_o)e^{\frac{\tau_N\tilde{\omega}}{\mu_o}} + \mu_o\right)^2},$$

where we define $\theta \equiv \mu_o/\tilde{\omega}$. Note that above derivative is positive by Lemma 1 since we assume that $\tau_N \geq 4$ and $\mu_o/\tilde{\omega} \geq 2$. Thus, we have that $\overline{\mathcal{G}}_{nc}$ is increasing in μ_o, which leads to the fact that

$$\mathcal{G}_{nc} \leq \lim_{\mu_o\to\infty}\overline{\mathcal{G}}_{nc} = \frac{1}{\tau_N - 1},$$

since $\lim\limits_{x\to\infty} x(e^{\tau/x} - 1) = \tau$.

B.3. Supplementary Results for the Proof of Proposition 3

LEMMA 1. *Let* $f(\tau,\theta) = e^{\tau/\theta}\left(\frac{\tau}{\theta^2} - \frac{\tau}{\theta} + 1\right)$*. Then,* $f(\tau,\theta) < 1$ *for all* $\tau \geq 4$ *and* $\theta \geq 2$*.*

Proof of Lemma 1 : By taking the derivative of $f(\tau,\theta)$ with respect to τ, we have that

$$\frac{\partial f(\tau,\theta)}{\partial\tau} = \frac{e^{\tau/\theta}(\theta(1-\tau)\tau + \theta)}{\theta^3} < \frac{e^{\tau/\theta}(2(1-\tau)+\tau)}{\theta^3} < 0$$

for all $\tau \geq 4$ and $\theta \geq 2$. As the above inequality proves the monotonicity of $f(\tau,\theta)$, we have that $f(\tau,\theta) < f(4,\theta)$ for all $\theta \geq 2$.

Furthermore, we have that

$$\frac{\partial f(4,\theta)}{\partial\theta} = \frac{8(\theta-2)e^{4/\theta}}{\theta^4} > 0,$$

for all $\theta \geq 2$, which implies that $f(4,\theta)$ is increasing in θ. Thus, we have that $f(\tau,\theta) < f(4,\theta) < \lim\limits_{\theta\to\infty} f(4,\theta)$. Finally, our claim holds because

$$\lim_{\theta\to\infty} f(4,\theta) = 1$$

Lemma 2. *Let* $g(\theta,x)=\frac{\theta^2(-x+e^x-1)}{e^x(\theta-1)-\theta}$*. Then,* $g(\theta,x)$ *is increasing in* x *for all* $x\geq\tau/\theta$*,* $\theta\geq 2$*, and* $\tau\geq 4$*.*

Proof of Lemma 2: By taking the derivative of $g(\theta,x)$ with respect to x, we have that

$$\frac{\partial g(\theta,x)}{\partial x}=\frac{\theta^2\left(e^x(\theta(x-1)-x)+\theta\right)}{\left(\theta\left(-e^x\right)+\theta+e^x\right)^2}.$$

Notice that the above derivative is positive when $h(\theta,x)\equiv e^x(\theta(x-1)-x)+\theta$ is positive. Hence, it is sufficient to show that $h(\theta,x)>0$ for all $x\geq\tau/\theta$, to prove our claim.

In order to show that $h(\theta,x)>0$ for all $x\geq\tau/\theta$, we first note that

$$h(\theta,\tau/\theta)=\theta\left[1-e^{\tau/\theta}\left(\frac{\tau}{\theta^2}-\frac{\tau}{\theta}+1\right)\right]>0,$$

where the inequality holds by Lemma 1. Furthermore, we have that

$$\frac{\partial h(\theta,x)}{\partial x}=e^x[(\theta-1)x-1]\geq e^x[(\theta-1)\tau/\theta-1]=e^x[\tau-\tau/\theta-1]\geq e^x[\tau/2-1]>0$$

for all $x\geq\tau/\theta$, $\tau\geq 4$, and $\theta\geq 2$. The above inequality implies that $h(\theta,x)$ is increasing in x for all $x\geq\tau/\theta$. Combining this with the fact that $h(\theta,\tau/\theta)>0$ proves that $h(\theta,x)>0$ for all $x\geq\tau/\theta$.

B.4. Proof of Theorem 2

We first want to note that

$$\begin{aligned}
\lim_{\mu_o\to\infty}[1-\tilde{\sigma}(K,\lambda_o,\mu_o)]\mu_o &= \frac{\tilde{\omega}e^{\zeta\tilde{\omega}}}{e^{\zeta\tilde{\omega}}-1}\\
\lim_{\mu_o\to\infty}\tilde{Q}(K,\lambda_o,\mu_o)/\mu_o &= \frac{e^{\zeta\tilde{\omega}}-1-\zeta\tilde{\omega}}{\tilde{\omega}(e^{\zeta\tilde{\omega}}-1)}\\
\lim_{\mu_o\to\infty}[1-\tilde{\sigma}_2(K,\lambda_o,\mu_o)]*\mu_o &= \frac{\tilde{\omega}\left(e^{\zeta\tilde{\omega}}+\gamma K-1\right)}{e^{\zeta\tilde{\omega}}-1}\\
\lim_{\mu_o\to\infty}\tilde{Q}_2(K,\lambda_o,\mu_o)/\mu_o &= \frac{e^{\zeta\tilde{\omega}}-1}{\tilde{\omega}\left(e^{\zeta\tilde{\omega}}+\gamma K-1\right)},
\end{aligned}$$

where $\zeta\equiv\lim_{\mu_o\to\infty}\tau_N/\mu_o$. Then, we use the above limits to construct an approximation for $\tilde{\Pi}_2(K,\lambda_o,\mu_o)$ such that

$$\lim_{\mu_o\to\infty}\tilde{\Pi}_1(K,\lambda_o,\mu_o)+\tilde{\Pi}_2(K,\lambda_o,\mu_o)=\lim_{\mu_o\to\infty}\hat{\Pi}_{sc}(K,\lambda_o,\mu_o),$$

where

$$\begin{aligned}
\hat{\Pi}_{sc}(K,\lambda_o,\mu_o)\equiv\ & V\mu_o\left(1-\frac{\tilde{\omega}e^{\tau_N\tilde{\omega}/\mu_o}}{\mu_o(e^{\tau_N\tilde{\omega}/\mu_o}-1)}-\frac{\tilde{\omega}\left(e^{\tau_N\tilde{\omega}/\mu_o}+\gamma K-1\right)}{\mu_o(e^{\tau_N\tilde{\omega}/\mu_o}-1)}\right)\\
& -h\mu_o\left(\frac{e^{\tau_N\tilde{\omega}/\mu_o}-1-\tau_N\tilde{\omega}/\mu_o}{\tilde{\omega}(e^{\tau_N\tilde{\omega}/\mu_o}-1)}+\frac{e^{\tau_N\tilde{\omega}/\mu_o}-1}{\tilde{\omega}\left(e^{\tau_N\tilde{\omega}/\mu_o}+\gamma K-1\right)}\right).
\end{aligned}$$

The above equality implies that $\lim\limits_{\mu_o\to\infty}\frac{K^*_{sc}}{\hat{K}_{sc}}=1$, where

$$\hat{K}_{sc}\equiv\arg\max_K\ \hat{\Pi}_2(K,\lambda_o,\mu_o).$$

Finally, by solving the above optimization problem, we explicitly obtain that

$$\hat{K}_{sc}=\frac{\left(e^{\tau_N\tilde{\omega}/\mu_o}-1\right)\left(\sqrt{hV\mu_o}-V\tilde{\omega}\right)}{\gamma V\tilde{\omega}}.$$

Notice that the above expression is an increasing function of h and τ_N.

B.5. Proof of Proposition 4

Using the limit results we have in the proof of Theorem 2, we have that

$$\lim_{\mu_o\to\infty} \tilde{\Pi}_1(K,\lambda_o,\mu_o) = \lim_{\mu_o\to\infty} \hat{\Pi}_{nc}(\lambda_o,\mu_o),$$

where

$$\hat{\Pi}_{nc}(\lambda_o,\mu_o) \equiv \mu_o \left[V\left(1-\frac{\tilde{\omega}e^{\tau_N\tilde{\omega}/\mu_o}}{\mu_o(e^{\tau_N\tilde{\omega}/\mu_o}-1)}\right) - h\left(\frac{e^{\tau_N\tilde{\omega}/\mu_o}-1-\tau_N\tilde{\omega}/\mu_o}{\tilde{\omega}(e^{\tau_N\tilde{\omega}/\mu_o}-1)}\right)\right]$$

Then, we can write the limiting behavior of the benefits from disguised queues as

$$\begin{aligned}\lim_{\mu_o\to\infty}\mathcal{G}_{sc} &= \lim_{\mu_o\to\infty}\frac{\tilde{\Pi}_1(K^*_{sc},\lambda_o,\mu_o)+\tilde{\Pi}_2(K^*_{sc},\lambda_o,\mu_o)}{2\tilde{\Pi}_1(0,\lambda_o,\mu_o)}-1\\ &= \lim_{\mu_o\to\infty}\frac{\hat{\Pi}_{sc}(\hat{K}_{sc},\lambda_o,\mu_o)}{2\hat{\Pi}_{nc}(\lambda_o,\mu_o)}-1 = \frac{1}{2\left[\frac{V\tilde{\omega}[e^{\zeta\tilde{\omega}}-1]}{h(e^{\zeta\tilde{\omega}}-\zeta\tilde{\omega}-1)}-1\right]}\end{aligned}$$

The above limit is increasing in h since $\frac{V\tilde{\omega}[e^{\zeta\tilde{\omega}}-1]}{h(e^{\zeta\tilde{\omega}}-\zeta\tilde{\omega}-1)}$ is decreasing in h. Similarly, The above limit is increasing in ζ since

$$\frac{d}{d\zeta}\left(\frac{V\tilde{\omega}[e^{\zeta\tilde{\omega}}-1]}{h(e^{\zeta\tilde{\omega}}-\zeta\tilde{\omega}-1)}\right) = -\frac{V\tilde{\omega}^2[e^{\zeta\tilde{\omega}}(\zeta\tilde{\omega}-1)+1]}{[h(e^{\zeta\tilde{\omega}}-\zeta\tilde{\omega}-1)]^2} < 0,$$

where the inequality holds by the fact that $e^x(x-1)+1>0$ for all $x\geq 0$.

To prove our second claim, we first note that

$$\begin{aligned}\lim_{\tau_N\to\infty}\tilde{\Pi}_1(K,\lambda_o,\mu_o) &= \mu_o\left(V-\frac{h}{\tilde{\omega}}\right)-V\tilde{\omega},\\ \lim_{\tau_N\to\infty}\tilde{\Pi}_2(K,\lambda_o,\mu_o) &= \frac{(\mu_o-\tilde{\omega})(\mu_oV\tilde{\omega}-h(\mu_o-\tilde{\omega}))}{\mu_o\tilde{\omega}},\end{aligned}$$

for all $K\leq 0$.

Then after some alegbra, we have that

$$\begin{aligned}\lim_{\tau_N\to\infty}\mathcal{G}_{sc} &= \lim_{\tau_N\to\infty}\frac{\tilde{\Pi}_1(K^*_{sc},\lambda_o,\mu_o)+\tilde{\Pi}_2(K^*_{sc},\lambda_o,\mu_o)}{2\tilde{\Pi}_1(0,\lambda_o,\mu_o)}-1\\ &= \frac{h\tilde{\omega}(2\mu_o-\tilde{\omega})}{2\mu_o(V\tilde{\omega}(\mu_o-\tilde{\omega})-h\mu_o)}.\end{aligned}$$

Finally, the above limit is positive because we must be in a parameter setting where $\tilde{\Pi}_1(K^*_{sc},\lambda_o,\mu_o)\geq 0$. Furthermore, have that

$$\lim_{\tau_N\to\infty}\tilde{\Pi}_2(K^*_{sc},\lambda_o,\mu_o)-\tilde{\Pi}_1(K^*_{sc},\lambda_o,\mu_o) = h\left(\mu_o-\frac{\tilde{\omega}}{\mu_o}\right)>0,$$

where the inequality holds since $\mu_o>\tilde{\omega}=\mu_o-\lambda_o$.