Managing Customer Expectations and Priorities in Service Systems

Qiuping Yu
Kelley School of Business, Indiana University, qiupyu@indiana.edu

Gad Allon
Kellogg School of Management, Northwestern University, g-allon@kellogg.northwestern.edu

Achal Bassamboo
Kellogg School of Management, Northwestern University, a-bassamboo@kellogg.northwestern.edu

Seyed Iravani
Industrial Engineering and Management Sciences, Northwestern University, s-iravani@northwestern.edu

We study how to use delay announcements to manage customer expectations while allowing the firm to prioritize among customers with different sensitivities to time and value. We examine this problem by developing a framework which characterizes the strategic interaction between the firm and heterogeneous customers. When the firm has information about the state of the system, yet lacks information on customer types, delay announcements play a dual role: they inform customers about the state of the system, while they also have the potential to elicit information on customer types based on their response to the announcements. The tension between these two goals has implications to the type of information that can be shared credibly.

To explore the value of the information on customer types, we also study a model where the firm can observe customer types. We show that having information on the customer type may improve or hurt the credibility of the firm. While the creation of credibility increases the firm’s profit, the loss of credibility does not necessarily hurt its profit.

Key words: delay announcements; heterogeneous customers; priority queue; information asymmetry; cheap talk

1. Introduction

Delay announcements are common practice in service systems, such as call centers, restaurants and airports. There is a variety of delay announcements used in practice. Some of the announcements provide little information, e.g., United Airline provides the announcement: “Due to high volume of calls, we are unable to answer your call immediately.” There are firms that provide fairly detailed announcements, e.g., ComEd provides the announcement: “your waiting time is about 4 minutes.” In service systems where the queue is not visible to customers, delay announcements may impact customers’ behavior. Consequently, in order to maximize the service provider’s value and mini-
mize the costs, it is important for it to understand how delay announcements influence customer behavior. However, this is a complex problem, which depends on the dynamics of the underlying queuing system, the structure of the delay announcements and customers’ strategic behavior. Moreover, in practice, the customer population is often heterogeneous along various dimensions and the firm may have limited capability to segment customers. In this paper, we study how to use delay announcements to manage customer expectations while allowing the firm to prioritize among heterogeneous customers in service systems.

Delay announcements have been studied in the literature. To put our work in perspective, we next briefly summarize previous work about delay announcements and we will do a detailed review in Section 2. Previous work assumes customers treat information as credible and implicitly assumes that the firm restricts the strategy to be truth-telling. Furthermore, these models assume that the firm’s strategy on providing information is a-priori fixed: the firms either give full information or no information and the information is quantifiable. The main issues with these assumptions are as follows: customers may not be naive and take the information for granted; the information might not be quantifiable: in the previous models, customers can take the information itself and compute their utilities. However, in many situations, the information needs further processing. For the information that the queue is long or that all agents are busy, customers can not simply convert such information into utilities that they obtain. In order to account for the factors mentioned above, we take a similar approach as the one used in Allon et al. (2011). However, it is important to note that Allon et al. (2011) focuses on the setting where customers are homogeneous. In our work, we allow customers to be heterogeneous. The fact that customers are heterogeneous brings three important features into our model: (1) customers may have private information about their types; (2) the firm may want to elicit information from customers about their types; and, (3) the firm may want to prioritize customers when necessary to maximize the profits. We examine the ability for the firm to sustain an equilibrium with influential cheap talk in such settings.

The goal of this work is to study how to use delay announcements to manage customers’ expectations and priorities in the presence of the heterogeneous customers. In order to do so, we consider a system with a single service provider. Customers arrive according to a Poisson process and the service time is exponentially distributed. Customers arrive to seek the rewards of service, while they incur costs due to waiting in the system. There are two types of customers, who differ in their rewards of being served and their waiting costs per unit time. As for the firm, it obtains values by serving customers and incurs costs for holding customers in the system. The value that the firm obtains by serving a customer is different for customers of different types. When customers arrive, the firm provides announcements to inform the customers about their anticipated delay. We start by focusing on the model where the firm does not observe the type of customers upon their
Title: Managing Customer Expectations and Priorities in Service Systems?

arrivals. Customers make decisions on whether or not to join the system based on the announcements received and their own types. Customers choose the actions to maximize their own utilities, while the firm decides what announcements to provide and prioritize customers appropriately when necessary to maximize its profits. In order to study the value that the firm may gain or lose by observing the type of customers, we also investigate a model where we allow the firm to observe customer types in Section 5.

One of the unique features of the model where the firm does not directly observe customer types is that both the customers and the firm have private information of their own: the customers have private information about their own types, while the firm has private information about the status of the system. To this end, as one may expect, the firm may need to elicit information from customers regarding their types and then prioritize them appropriately if necessary to maximize the profits. However, our results show that it may not be necessary for the firm to fully differentiate customers of different types to achieve the first best, when the per unit holding cost is the same for all customers. Partially separating customers could be sufficient to achieve the first best solution. In particular, we show that the optimal announcement policy used by the firm to achieve the first best can be characterized with two different thresholds. Hence, the firm only needs three signals to indicate the number of customers in the system relative to the two thresholds to achieve the first best. Furthermore, we show that under certain conditions, a pooling equilibrium, where the firm does not elicit information on customer types at all, may perform the best in firm’s profit among all equilibria. The announcement policy in the pooling equilibrium can be characterized by one single threshold. In particular, the firm provides two different signals to indicate whether the number of customers in the system is below or above the threshold.

Our results above show that the firm can achieve the first best when the per unit holding costs are the same for all customers by only partially separating the customers. However, we find that, when the per unit holding costs are different for customers of different types, the firm cannot achieve the first best through delay announcements. Note that the optimal policy of the firm in this case is comprised of two components. In particular, other than providing delay announcements to induce the desired customer behavior in terms of whether to join the system, the firm may also want to prioritize the customers who joined the system based on their types to minimize the costs. However, the firm can only elicit information on customers types when customers respond to the announcements differently. Moreover, one can show that there is at least one message which induces both customer types to join. Thus, the firm cannot fully separate the customers by using only delay announcements, which prevents the firm from achieving the first best when the per unit holding costs are different for customers of different types.
The question now is how the firm can influence the customers through delay announcements to maximize profits, when the holding costs are different for customers of different types. Recall that the optimal strategy of the firm includes two components: the priority policy and the announcement policy. We start by exploring the optimal priority policy. Given the firm does not directly observe customer types, it can only prioritize the customers in the system based on the announcements that they receive. In particular, we show that it is optimal for the firm to give absolute priority to customers in the system who receive the announcements corresponding to the smaller expected per unit holding cost. As for the firm’s announcement policy, our results show that the firm cannot improve its profits by using more than three announcement. However, we are not able to characterize the structure of the optimal announcement policy in details when the per unit holding costs are different for customers of different types.

So far, we focus on the influential equilibrium where the firm provides creditable information. In practice, however, there are many service providers that share no information whatsoever with the customers or information uncorrelated with the state of the system. To this end, we explore whether these systems are in equilibrium. We show that an equilibrium where no meaningful information provided by the firm and customers disregard the announcements may indeed exist. We refer to such an equilibrium as a babbling equilibrium (Allon et al. (2011)). We find that the firm always prefer the influential equilibria to the babbling equilibria. Meanwhile, from customer’s perspective, while customers are better off in the influential equilibria in terms of their overall utility, the utility of the more patient customer type may get hurt. This is consistent with the empirical results in Yu et al. (2015). One possible explanation is that more customers of the less patient type would join the system when they are provided with announcements, which may hurt the utility of the more patient customers.

Recall that we assume that the firm does not observe the types of customers in the model above. However, in practice, many firms have information on customer types. For example, the call center we worked with for our empirical paper, see Yu et al. (2015), has very detailed information about the majority of their customers. To study the value that the firm may gain or lose when the firm has information about customer types, we next extend the model above to allow the firm to observe the types of customers upon their arrival. We show that, in any equilibria with influential cheap talk, the firm achieves the first best when it can observe customer types. In order to achieve the first best, the firm gives absolute priority to the type of customers who have a higher per unit holding cost between the two customer types. As for the optimal announcement policy, we show that it can be characterized by two monotonic decreasing switching curves. Moreover, we find that information on customer types may extend the region where the firm can achieve influential equilibria. The intuition is that, when the firm observes the type of the customer, the firm can
provide announcements to customers based on their types to better match their expectations. However, information on customer types may also detract from the resulting influential equilibrium. This occurs because, when the firm has information on customer types, it attempts to generate more profits from the customers which leads to more misalignment between the firm and the customers. We find that the creation of credibility in the expansion region improves the firm’s profit. Similarly, one may expect the loss of credibility in the contraction region to hurt the firm’s profit. However, we show that the loss of credibility may even improve the firm’s profits.

2. Literature Review

As we study the use of delay announcements to manage customers, we divide the relevant literature into the following branches: queuing models with delay announcements, admission control, pricing in priority queue, and cheap talk games.

Queuing Models with Delay Announcements. One of the first papers that discusses the question of whether to reveal the queue length information to customers is Hassin (1986), which studies the problem of whether a price-setting, and revenue-maximizing service provider should provide the queue length information to arriving customers when it has the option to do so. It is shown that it may, but not always be socially optimal to provide the queue length information, and that it is never optimal to encourage suppression when the revenue maximizer prefers to reveal the queue length. Whitt (1999) brings the concept of information revelation to the specific setting of call centers, where call centers communicate with their customers about the anticipated delay by providing delay announcements. The author studies the impact of informing customers about anticipated delay in a single class Markovian call center model, and shows that average waiting time can be reduced when accurate announcements are provided. Guo and Zipkin (2007) extends the model above by studying the impact of delay announcements with different information accuracy. They show that accurate delay information might improve or hurt the system performance.

While all these papers assume customers do not abandon the system once they join the queue, Armony et al. (2009) relaxes this assumption. Armony et al. (2009) studies the performance impact of making delay announcements to arriving customers in a many-server queue setting with customer abandonment. Customers who must wait are provided with either the delay of the last customer to enter service or an appropriate average delay upon arrival. The authors show that within the fluid-model framework, under certain conditions, the actual delay coincides with the announced delay. Motivated by this type of delay announcements, Ibrahim and Whitt (2009) explores the performance of different real-time delay estimators based on recent delay experienced by customers, allowing for customer abandonment.

All the aforementioned works assume that the information is credible and is treated as such by customers. To this end, it is important to note that Yu et al. (2015) has provided empirical evidence
indicating that customers may be able to strategically interpret the announcement. In particular, the authors show that the structural model where customers match the delay announcement with the actual offered waiting time associated with it better explains the observed customer behavior than the model where customers make abandonment decisions based on the explicit delay estimates provided in the announcements. Allon et al. (2011) has accounted for such strategic customer. Specifically, the authors exam the problem of information communication by considering a model in which both the firm and the customers act strategically: the firm in choosing its delay announcement while anticipating customer response, and the customers in interpreting these announcements and in making the decision on whether to join the system. Our paper extends the model in Allon et al. (2011) by incorporating the heterogeneity of the customer population, while allowing the firm to prioritize customers when necessary to maximize profits.

**Admission Control.** In our model, customers terminate their calls or request for service based on their assessment of the service quality and the firm does not have control over customers’ decisions. However, as we will show in the paper, the firm may achieve the first best solution through delay announcements as if it had full control over customers’ admission. To this end, our paper is related to the literature of admission control, which starts from Naor (1969). The author shows customers are more patient than what a social planner would like them to be. The imposition of tolls may lead to attainment of social optimality. Rue and Rosenshine (1981) extends the model above to the setting with multiple customer classes who are first-come, first-served. Similarly, the authors show that both the individual and social optimal policies are threshold-type policies. In addition, the social optimal threshold is shown to be no greater than the individual one for the same customer class. While none of the works mentioned above consider service priorities, Chen and Kulkarni (2007) takes one step further and studies the admission control problem for queuing system serving two customer classes with priority. Class 1 customers have preemptive resume priority over class 2 customers, while customers from the same class are served in a first-come, first-served basis. The authors show that optimal policy is of either threshold-number or switching-curve form under individual, social or class-specific optimization criterion. Instead of focusing on the queuing systems, Iravani et al. (2012) investigates optimal production and admission control policies in manufacturing systems that produce two types of products: one type consists of identical items that are produced to stock, while the other has varying features and is produced to order. The authors characterize the optimal production and admission policies with a partial-linear structure. Moreover, they provide insights about the benefits of the new policies using computational analysis.

**Pricing in Priority Queue.** In the presence of multiple customer classes and when the firm does not observe customer types or does not have direct control on customers’ priorities, pricing is one of the commonly used tools to differentiate customers and then prioritize them when necessary.
Mendelson and Whang (1990) suggests a pricing mechanism to optimize the overall social welfare in an $M/M/1$ system with multiple types of customers. In particular, the paper shows that the $c\mu$ rule is incentive compatible to optimize the social welfare. Furthermore, the price charged should be equal to the expected externalities that the customer imposes on the system by joining the priority group conditioning on her service requirements. Afeche (2004) extends the model in Mendelson and Whang (1990) to study how the firm should design a incentive compatible pricing-scheduling mechanism to maximize its revenue, given that there are two types of customers. The paper shows that the $c\mu$ rule need not be optimal and it may be necessary to add a strategic delay to achieve the optimal in certain settings. The papers above show that one may design a direct revelation mechanism to achieve the optimal result with pricing strategies. However, there are organizations where pricing strategies are not preferred or allowed, e.g., Disneyland, DMV or IRS. To address the problems that arise in these contexts, our paper aims to explore how to manage customer expectations and priorities using delay announcements.

**Cheap Talk Game.** The framework used in this paper is inspired by the classical cheap talk model proposed in Crawford and Sobel (1982). The authors introduced a cheap talk game model of strategic communication between a sender and a receiver. In this model, the sender, who has private information, sends possibly noisy information to the receiver, who then takes payoff-relevant actions. It’s important to note that the distribution of the sender’s private information is given exogenously and does not depend on the equilibria of the game. However, in our endogenous cheap talk setting, the distribution of the private information depends on the equilibrium of the game. Driven by the specific queuing application, our model has two novel features: first, the game is played with multiple and different types of receivers (customers) whose actions have externalities on other receivers; and second, the stochasticity of the state of the system is not exogenously given but is determined endogenously. In particular, the private information in this model (i.e., the queue length) is driven by the system dynamics, which in turn depends on the equilibrium strategies of both the firm and the customers. As we shall see, the multiplicity of receivers with externalities and the endogenous uncertainty impact both the nature of the communication and the outcome for the various players. Hence, while the framework used in this paper echoes the cheap-talk model described in the literature, the above mentioned distinguishing features lead to different results.

Allon et al. (2011) appears to be the first paper in the operations management literature to consider a model in which a firm provides unverifiable real time dynamic delay information to its customers. As we mentioned earlier Allon et al. (2011) focuses on the scenario where customers are homogeneous, while this paper accounts for the heterogeneity of the customers. The fact that customers are heterogeneous brings three unique features to our model: 1) both the customers and the firm have private information of their own; 2) the firm may tend to elicit information on
customers types through delay announcements; and 3) the firm may want to prioritize customers of different types.

Organization of the remainder of the paper: Section 3 provides the detailed description of the model where the firm has no information on customer types and our notion of equilibrium. In Section 4, we state our main results for the equilibria in the model with no information on customers types. Section 5 explores the value that the firm and the customers may gain or lose when the firm observes the type of customers compared to the case when the firm does not. In Section 6, we provide concluding remarks. All proofs are relegated to the Appendix B.

3. Model with No Information on Customer Types

We consider a service system with a single service provider, where customers arrive according to a Poisson process with rate $\lambda$ and the service time is exponentially distributed with rate $\mu$. We assume that there are two types of customers, which we refer to as low and high type customers denoted by $L$ and $H$, respectively. With probability $\beta_i$, for $i \in \{H, L\}$, an arriving customer is a type $i$ customer. Customers arrive to seek service and get rewards from the service, while they incur costs due to their waiting in the system. The reward of being served for type $i$ customers is denoted as $R_i$, while the waiting cost per unit time is denoted as $c_i$, for $i \in \{H, L\}$. From the firm’s perspective, the firm obtains value from serving customers, while it incurs costs for holding customers in the system. Let us denote the value that the firm obtains from serving a type $i$ customer as $v_i > 0$, for $i \in \{H, L\}$. Without loss of generality, we let $v_H > v_L$. As for the holding costs that the firm incurs, they include, among others, the goodwill cost due to the long wait, the cost of providing the actual waiting space and facilities, the opportunity cost of missing the chances to have customers generating revenues at some other facilities within the firm. In particular, we denote the per unit time holding cost of a type $i \in \{H, L\}$ customer as $h_i$. We assume all the above parameters are known to both the customers and the firm. When customers arrive, the firm provides delay announcements to customers possibly based on the current congestion in the system. We focus on the scenario where the firm cannot observe customer types before it provides announcements in this section. We will relax this assumption in Section 5 in order to explore the value that the firm may gain or lose by observing customer types. Based on the announcements received, customers make decisions on whether to join the system by trading off between their rewards of being served and their waiting costs. To characterize the interactions between customers and the firm through delay announcements, we next define the game that both the customers and the firm engage in.

The utility of a type $i$ customer, for $i \in \{H, L\}$, is given by

$$u_i(a_i, w) = \begin{cases} R_i - c_i w & \text{if } a_i = \text{join} \\ 0 & \text{if } a_i = \text{balk} \end{cases}$$

(1)
where $a_i$ is type $i$ customers’ decision on whether to join the system and $w$ is customers’ waiting time in the system. Note that to maximize utility, customers of type $i$, $i \in \{H, L\}$, would like to join the system when the expected waiting time in the system is smaller than $\frac{R_i}{c_i}$, balk otherwise. To this end, we refer $\frac{R_i}{c_i}$ as the patience of type $i$ customers with $i \in \{H, L\}$. Throughout the paper, we assume that $\frac{R_i}{c_i} > \frac{1}{\mu}$, for $i \in \{H, L\}$, so that customers of both types are better off joining the system when there is no waiting. Otherwise, customers would not join the system even when there is no delay, and it is not necessary to provide delay announcements at all. Meanwhile, the firm’s profit by serving a customer of type $i$, for $i \in \{H, L\}$, is given by $v_i - h_i w$. We assume $h_i > 0$ for all $i \in \{H, L\}$, so that the firm would have incentive not to admit either customer types beyond certain finite threshold.

In our model, we assume customer types are private information of the customers, while the current state of the system, i.e., the number of customers in the system, is private information of the firm. To investigate how delay announcements impact customers’ behavior and what announcements the firm should provide to maximize its profits, we next formally describe the game played between the firm and the customers. The equilibrium concept that we use is a Markov Perfect Bayesian Nash Equilibrium (MPBNE). In our case, it is simply a set of strategies of the firm and the customers at Nash Equilibrium that describes how customers incorporate delay announcements and their own types to their decisions on whether to join the system, and how the firm chooses announcements to maximize its profits. MPBNE only allows actions to depend on pay-off relevant information, which rules out strategies that depend on non-substantive moves by the opponent. We will formally define MPBNE later in this section.

To describe the announcements, let $M = \{m_1, m_2, m_3...\}$ be the set of possible discrete messages provided by the firm. The messages could be quantitative or qualitative. For example, Apple stores provide queue position information to customers, where the possible messages used are any non-negative integers; while Citibank provides information: “all agents are currently serving other customers, please hold,” which is qualitative. To characterize the interaction between the customers and the firm, we start from how customers respond to announcements. Once customers receive announcements from the firm, they decide whether to join the system based on the messages received and their own types. Customer of different types may respond differently to the same announcement due to different waiting costs and rewards that they receive from being served. In particular, customers’ action rule is given by a function $a_i : M \rightarrow \{1, 0\}$, for $i \in \{H, L\}$. Moreover, $a_i(m) = 1$ means the type $i$ customer joins the system when she receives the message $m$, while $a_i(m) = 0$ represents that she balks.

We next turn to define the strategy of the firm. Note that the firm’s optimal strategy is comprised of two components in our model: 1) the firm decides what announcements to provide based on the
number of customers from each type to induce desired customer response, and 2) given that there are two types of customers in the system and the firm cannot directly observe customer types, the firm may want to elicit information on customer types and prioritize them when necessary.

Let us start from the announcement policy. To make a better decision on what announcements to provide, the firm may want to elicit as much information on customer types as possible. However, the firm can only differentiate customers, when they respond to announcements differently. Thus, instead of differentiating customers based on their types, the firm can only classify customers based on the announcements that they receive. According to the action rule defined above, there are two different reactions, i.e., join and balk, for each customer type. Thus, we can classify the announcements into four categories given there are two types of customers. In particular, the first category includes announcements under which both customer types balk. The second category includes announcements under which only the high type customers join the system but not the low type. The third category includes announcements under which only the low type customers join the system but not the high type, while the fourth category includes the announcements under which both customer types join the system. To represent these four categories of announcements, we let $M_O$ be the set of announcement where customers of type $i \in O$ join and customers of type $i \in O^c$ balk. Thus, we have $M_\emptyset$, $M_{\{H\}}$, $M_{\{L\}}$ and $M_{\{H,L\}}$ denoting the four categories of announcement sets mentioned above, respectively. One can see that $M_\emptyset$, $M_{\{H\}}$, $M_{\{L\}}$ and $M_{\{H,L\}}$ are all subsets of $M$, which is the set of the possible messages provided by the firm. Moreover, the message subsets $M_\emptyset$, $M_{\{H\}}$, $M_{\{L\}}$ and $M_{\{H,L\}}$ are mutually exclusive. To this end, the firm can classify the customers in the system into three categories: customers receiving announcements from $M_{\{H\}}$, $M_{\{L\}}$ or $M_{\{H,L\}}$. In particular, we let $n_H$, $n_L$ and $n_{HL}$ denote the number of customers in the system that received announcements from the subsets $M_{\{H\}}$, $M_{\{L\}}$ and $M_{\{H,L\}}$, respectively.

We are now ready to formally define the announcement policy of the firm. In particular, the announcement policy of the firm can be characterized by a function $A : S \mapsto M$, where $S$ is the set of system states with $S = \{(n_H, n_L, n_{HL}) | (n_H, n_L, n_{HL}) \in \mathbb{Z}^3\}$. For example, we have $A(n_H, n_L, n_{HL}) = m$, if the firm provides the announcement $m$ to the next arriving customer when there are $n_H$, $n_L$ and $n_{HL}$ customers in the system who, upon arrival, received announcements from the message subsets $M_{\{H\}}$, $M_{\{L\}}$ and $M_{\{H,L\}}$, respectively. Similarly, the scheduling policy of the firm is a function which maps the current system state to the next customer to serve. As we mentioned earlier, the firm can only distinguish the customers based on the announcements they receive. In particular, the firm can sort the customers in the system into three categories: customers receiving announcements from $M_{\{H\}}$, $M_{\{L\}}$ or $M_{\{H,L\}}$. To this end, the firm can schedule the customers based on the announcements. In particular, the firm’s scheduling policy is then given by a function $g : S \mapsto X$, where $S$ is the set of system states and $X$ is the set of announcement types with
$X = \{ M_0, M_{(H)}, M_{(L)}, M_{(H,L)} \}$. For example, we have $g(n_H, n_L, n_{HL}) = M_{(L)}$, if the next customer to serve is the first customer in the system who received an announcement from the message subset $M_{(L)}$, when there are $n_H$, $n_L$ and $n_{HL}$ customers in the system receiving announcements from the message subsets $M_{(H)}$, $M_{(L)}$ and $M_{(H,L)}$, respectively.\(^1\)

Note that the steady-state probability distribution of the system state $( n_H, n_L, n_{HL} )$ depends on the both the customer strategy, $a_i$, with $i \in \{ H, L \}$, the firm’s scheduling $g$ and announcement policy $A$. Let $p(n_H, n_L, n_{HL}|a, g, A)$ represent the steady-state probability of state $( n_H, n_L, n_{HL} )$, conditional on the type $i$ customers’ strategy $a_i$, the firm’s announcement policy $A$ and scheduling policy $g$ with $i \in \{ H, L \}$. Meanwhile, we let $w^g_m(n_H, n_L, n_{HL})$ denote the waiting time of the customer who receives the announcement $m$ and joins the system at state $( n_H, n_L, n_{HL} )$.

Recall that the equilibrium concept we employ is MPBNE. We now formally describe the pure strategy MPBNE in the following definition.

**Definition 1 (Markov Perfect Bayesian Nash Equilibrium).** We say that the firm’s announcement policy $A(\cdot)$, scheduling policy $g(\cdot)$ and customers’ action rule $a_i(\cdot)$ with $i \in \{ H, L \}$, form a Markov Perfect Bayesian Nash Equilibrium (MPBNE), if they satisfy the following conditions:

1. For each $m \in M$ and $i \in \{ H, L \}$, we have

   $$a_i(m) = \begin{cases} 
   1 & \text{if } \frac{\Sigma_{(a_H, n_L, n_{HL})} \cdot \lambda \cdot p(n_H, n_L, n_{HL}|a, g, A)}{\Sigma_{(a_H, n_L, n_{HL})} \cdot \lambda \cdot p(n_H, n_L, n_{HL}|a, g, A)} \geq 0 \\
   0 & \text{otherwise,}
   \end{cases} \quad (2)$$

2. There exists value functions $v(n_H, n_L, n_{HL})$ with $( n_H, n_L, n_{HL} ) \in \mathbb{Z}^3$, constant $\gamma$, and the announcement policy $m = A^*(n_H, n_L, n_{HL})$ that solve the following equation:

   $$V(n_H, n_L, n_{HL}) = \frac{1}{\Lambda} \left\{ - (h_H \beta_H + h_L \beta_L) n_{HL} - h_L n_L - h_H n_H 
   \right.$$

   $$\left. + \lambda \max_{m \in M} \left\{ (V(n_H, n_L, n_{HL} + 1) + \beta_H v_H + \beta_L v_L) a_H(m) a_L(m) 
   \right. 
   \left. + (\beta_H V(n_H + 1, n_L, n_{HL}) + \beta_L V(n_H, n_L, n_{HL}) + \beta_H v_H) a_H(m)(1 - a_L(m)) 
   \right. 
   \left. + (\beta_H V(n_H, n_L, n_{HL} + 1) + \beta_L V(n_H, n_L + 1, n_{HL}) + \beta_L v_L) a_L(m)(1 - a_H(m)) 
   \right. 
   \left. + V(n_H, n_L, n_{HL})(1 - a_H(m))(1 - a_L(m)) \right\} 
   \right.$$

   $$\mu \max \left\{ V(n_H - 1, n_L, n_{HL}) I_{\{n_H > 0\}} + V(n_H, n_L, n_{HL}) I_{\{n_H = 0\}}, 
   \right.$$

   $$\left. V(n_H, n_L - 1, n_{HL}) I_{\{n_L > 0\}} + V(n_H, n_L, n_{HL}) I_{\{n_L = 0\}}, 
   \right.$$

   $$\left. V(n_H, n_L, n_{HL} - 1) I_{\{n_{HL} > 0\}} + V(n_H, n_L, n_{HL}) I_{\{n_{HL} = 0\}}, 
   \right.$$

   $$\left. V(n_H, n_L, n_{HL}) \right\} , \quad (3)$$

with $\Lambda = \lambda + \mu$. Thus, the firm’s optimal announcement policy is given by $m = A^*(n_H, n_L, n_{HL})$.

\(^1\) We have $g(n_H, n_L, n_{HL}) = M_0$, if the firm stays idle at state $( n_H, n_L, n_{HL} )$. 

---

**Title:** Managing Customer Expectations and Priorities in Service Systems?
3. The firm’s optimal scheduling policy is given as follows:

\[
g(n_H, n_L, n_{HL}) = \begin{cases} 
M_{(H)} & \text{if } V_1 = \max\{V_1, V_2, V_3, V_4\} \\
M_{(L)} & \text{if } V_2 = \max\{V_1, V_2, V_3, V_4\} \\
M_{(H,L)} & \text{if } V_3 = \max\{V_1, V_2, V_3, V_4\} \\
M_{\emptyset} & \text{if } V_4 = \max\{V_1, V_2, V_3, V_4\}, 
\end{cases}
\]

(4)

where we have

\[
V_1 = V(n_H - 1, n_L, n_{HL})I_{(n_H > 0)} + V(n_H, n_L, n_{HL})I_{(n_H = 0)},
\]

\[
V_2 = V(n_H, n_L - 1, n_{HL})I_{(n_L > 0)} + V(n_H, n_L, n_{HL})I_{(n_L = 0)},
\]

\[
V_3 = V(n_H, n_L, n_{HL} - 1)I_{(n_{HL} > 0)} + V(n_H, n_L, n_{HL})I_{(n_{HL} = 0)}
\]

and

\[
V_4 = V(n_H, n_L, n_{HL}).
\]

In the above MPBNE definition, the first condition, given by (2), describes the customers’ decision rule. In particular, customers join the system if the expected utility conditional on the firm’s announcement policy and the messages received is positive, and balk otherwise. The second condition, see (3), claims that the composite functions \(a_i \circ A\), for \(i \in \{H, L\}\), solve the firm’s Markov Decision Process (MDP) which is closely related to the admission control problem in the MDP literature, see Stidham (1985). The constant \(\gamma\) is the firm’s long-run average profit per unit time under the optimal policy, and the functions \(V(n_H, n_L, n_{HL})\) with \((n_H, n_L, n_{HL}) \in \mathbb{Z}^3\) are the relative profit of the firm for states \((n_H, n_L, n_{HL})\). Lastly, the last condition given by (4) characterizes the optimal scheduling policy of the firm. Note that we assume that the system parameters are known to both the customers and the firm in the model only for technical convenience. In fact, the insights throughout the paper will continue to hold if customers are able to form the correct belief about the offered waiting time associated with the delay announcements given the equilibrium strategies. To this end, it is important to note that Yu et al. (2015) provides strong empirical evidences supporting such an assumption. For simplicity, we focus on pure strategy equilibria throughout the paper.

4. Results: Model with No Information on Customer Types

Recall that the goal of this paper is to study how to manage customer expectations and priorities to maximize the firm’s profits. Specifically, we focus on delay announcements as the only mechanism available to the firm and study its opportunities and limitations. In order to do so, we start by showing the existence of equilibria and characterizing these equilibria where the firm provides credible delay announcements to induce the desired responses from the heterogeneous customers.
We will then explore the equilibria that emerge between the customers and the firm when the firm provides no announcements or announcements which are uncorrelated with the system states. We refer to such equilibria as babbling equilibria. Lastly, by comparing the babbling equilibria and the equilibria where the firm provides credible information, we provide insights on whether and how the firm should provide announcements to maximize profits. We will also discuss whether providing announcements benefits customers and the firm.

To characterize the equilibria that emerge between the customers and the firm, the first question that comes up is whether announcements impact customer behavior. To address this question, we introduce the following definitions.

**Definition 2 (Influential and Non-influential Equilibrium).**
1. We say that an MPBNE \((a_L, a_H, A, g)\) is *influential* if, \(\forall i \in \{H, L\}\), there exists two announcements \(m_1^i \) and \(m_2^i \) which are used with positive probability\(^2\) in the equilibrium so that we have \(a_i(m_1^i) \neq a_i(m_2^i)\).
2. We say that an MPBNE \((a_L, a_H, A, g)\) is *non-influential*, if we have \(a_i(m_1) = a_i(m_2)\), \(\forall m_1, m_2 \in M\) and \(i \in \{H, L\}\).

We say that the firm can credibly communicate with the customers through delay announcements if there exists an equilibrium with influential cheap talk. Given that there are two types of customers in the system, the next question is whether the announcements influence customers of different types differently. In order to address this question, we introduce the following definition.

**Definition 3 (Pooling, Semi-separating and Separating Equilibrium).**
1. We say that an MPBNE \((a_L, a_H, A, g)\) is a *pooling equilibrium* if, \(\forall m \in M\), which are used with positive probability in equilibrium, we have \(a_L(m) = a_H(m)\).
2. We say that an MPBNE \((a_L, a_H, A, g)\) is a *semi-separating equilibrium*, if \(\exists i, j \in \{H, L\}\) with \(i \neq j\), \(\forall m \in M\) that is used with positive probability in equilibrium, we have \(a_i(m) \geq a_j(m)\); moreover, there exists at least one message \(\hat{m} \in M\) which is used with positive probability in equilibrium, such that \(a_i(\hat{m}) > a_j(\hat{m})\) holds.
3. We say that an MPBNE \((a_L, a_H, A, g)\) is a *separating equilibrium* if \(\exists m_1, m_2 \in M\) and \(m_1 \neq m_2\), which are used with positive probability in equilibrium, such that \(a_L(m_1) > a_H(m_1)\) and \(a_L(m_2) < a_H(m_2)\) both hold.

Following the above definition, we refer to an influential equilibrium, where any given announcement influences customers of different types identically, as a *pooling equilibrium*. We refer to an influential equilibrium as a *separating equilibrium*, if there exists one announcement that induces low type customers to join and high type customers to balk, while another announcement that induces the exact opposite reactions from these two types of customers. Moreover, we refer to

\(^2\)We say that a message \(m\) is used with positive probability under an equilibrium \((a_L, a_H, A, g)\), if \(\sum_{(n_H, n_L, n_{HL}) \in \mathcal{N}_{a_L, a_H, A}} p(n_H, n_L, n_{HL} | a_L, a_H, A, g) > 0\).
an influential equilibrium between a pooling equilibrium and a separating equilibrium as a semi-separating equilibrium. From the cheap talk literature, one may expect that the equilibrium in cheap talk games is not unique even when it exists. It is because, one can always relabel the messages to induce other equilibria with the same outcomes and pay-offs for the firm and the customers. Similar to Allon et al. (2011), we introduce the definition for MPBNE being Dynamics and Outcome Equivalent (DOE) as follows.

**Definition 4 (Dynamic and Outcome Equivalent (DOE)).** We say that two MPBNE $(a_{1}^{L}, a_{1}^{H}, A_{1}, g_{1})$ and $(a_{2}^{L}, a_{2}^{H}, A_{2}, g_{2})$ are DOE, if $a_{i}^{1}(A_{1}(n_{H}, n_{L}, n_{HL})) = a_{i}^{2}(A_{2}(n_{H}, n_{L}, n_{HL}))$, $\forall i \in \{H, L\}$ and $\forall (n_{H}, n_{L}, n_{HL}) \in Z_{3}$.

It is important to note that the utility of each customer type and the profit of the firm are identical under any two MPBNEs which are Dynamics and Outcome Equivalent.

Recall that there are two different actions, i.e., join and balk, for each customer type. Thus, there are four possible customer reactions when there are two customer types: both customer types join the system, only the high type customers join the system, only the low type customers join the system, and both customer types balk. One may expect the firm to use four different announcements to induce the desired customer reactions in equilibria. However, the following proposition shows that, for any given pure strategy equilibrium, we can find a pure strategy equilibrium where the firm uses at most three announcements which is DOE to the given equilibrium. The main reason is that the second and the third reactions mentioned above, i.e., only the high type customers join the system, and only the low type customers join the system, are mutually exclusive in equilibria. The rigorous proofs of all results including this proposition are relegated to Appendix B.

**Theorem 1.** Given any pure strategy MPBNE for the two-class cheap talk game, there exists a pure strategy MPBNE which is DOE to the given equilibrium and in which the firm uses at most three announcements.

Following the theorem above, we can focus on the pure strategy equilibria where the firm uses at most three announcements without loss of generality.

### 4.1. Full Information and Full Control Solution

To construct the equilibria for the above model with no information on customer types, we start from a benchmark case where the firm not only has full control over customer admission, but also has full information on their types. We refer to the solution to this problem as the first best solution. Note that the firm’s first best solution is comprised of two components: the firm’s optimal admission policy and the its optimal scheduling policy. In particular, one shall see that the firm’s optimal admission policy may depend on the system states. When the firm observes the types
of the customers upon their arrivals, the system states can be characterized by the number of customers of each type. To characterize the system state, we let \( n_H^0 \) and \( n_L^0 \) be the number of high and low type customers in the system, respectively. Thus, the total number of customers in the system is given by \( n = n_H^0 + n_L^0 \). Moreover, we let \( S_t \) be the set of the system states when the firm observes the type of the customers. In particular, the set of the system states is given by \( S_t = \{(n_H^0, n_L^0)| (n_H^0, n_L^0) \in \mathbb{Z}^2\} \). Other than the admission policy, the firm may also want to schedule customers appropriately to optimize profits. The first two results in the following proposition show that the firm’s optimal admission policy can be characterized by two monotonically non-increasing switching curves. Furthermore, the last result of the proposition characterizes the firm’s optimal scheduling policy. In particular, we find that, when we have \( h_L \neq h_H \), it is optimal for the firm to give absolute priority to customers with a higher per unit holding cost. This shows that the \( c\mu \) rule, which was first established in Smith (1956), continues to hold in our setting.

**Lemma 1.** The first best solution of the firm is characterized as follows:

1. For each \( n_L^0 \geq 0 \), there exists a threshold \( S_H(n_L^0) \), such that a high type customer is accepted if and only if \( n_H^0 \leq S_H(n_L^0) \). Furthermore, \( S_H(n_L^0) \) is monotonically non-increasing in \( n_L^0 \).

2. For each \( n_H^0 \geq 0 \), there exists a threshold \( S_L(n_H^0) \), such that a low type customer is accepted if and only if \( n_L^0 \leq S_L(n_H^0) \). Furthermore, \( S_L(n_H^0) \) is monotonically non-increasing in \( n_H^0 \).

3. When we have \( h_L \neq h_H \), the firm gives preemptive resume priority to customers of type \( k \) in the system, where \( k \) is given by \( k = \arg \max_{i \in \{H, L\}} \{h_i\} \). When we have \( h_H = h_L \), the order of service does not impact the profit of the firm.

It is worth mentioning that, if there exists \( i \in \{H, L\} \) so that we have \( S_i(0) < 0 \), to achieve the first best solution, the firm will not admit type \( i \) customers regardless of the number of customers in the system. In this case, the system dynamics will be identical to the one discussed in Allon et al. (2011) where there is only one customer class. To this end, throughout this paper, we focus on the cases with \( S_H(0) \geq 0 \) and \( S_L(0) \geq 0 \).

Note that, when the per unit holding cost is the same for all customers, we can simplify the optimal admission policy of the firm characterized in the above proposition. In particular, when we have \( h_H = h_L \), we specify the two switching curves mentioned above in the following proposition.

---

3 Recall that, in the model with no information described in Section 3, the firm does not observe customer types and can only differentiate customers based on the announcements that they receive. To this end, the system states are characterized by the number of customers receiving each type of the announcements. In particular, the set of the system states \( S \) is given by \( S = \{(n_H, n_L, n_{HL})|(n_H, n_L, n_{HL}) \in \mathbb{Z}^3\} \), where \( n_H, n_L \) and \( n_{HL} \) are the number of customers in the system receiving announcements from the announcement sets \( M_H \), \( M_L \) and \( M_{HL} \), respectively. The total number of customers in the system is given by \( n = n_H + n_L + n_{HL} \).
Lemma 2. When $h_H = h_L$, the two switching curves given in Lemma 1, i.e., $S_L(n^0_H)$ and $S_H(n^0_L)$, are given by the following equations:

\[ S_L(n^0_H) = \hat{n}_L - n^0_H \quad \text{and} \quad S_H(n^0_L) = \hat{n}_H - n^0_L. \]

Moreover, $\hat{n}_L$ and $\hat{n}_H$ are two finite constants with $\hat{n}_L \leq \hat{n}_H$. These two constants are independent of the system state given by $(n^0_H, n^0_L)$.

Following Lemma 1, one can see that, when the per unit holding cost is different for customers of different types, the firm’s admission policy depends on both the number of low type customers in the system and the number of high type customers. Meanwhile, when the per unit holding cost is the same for all customers, the firm’s optimal admission policy only depends on the total number of customers in the system but not the types of the customers, see Lemma 2. In particular, when we have $h_H = h_L$, the firm’s optimal admission policy is given as follows: the firm accepts both low and high type customers if the total number of customers in the system is below $\hat{n}_L$; the firm accepts only high type customers if the total number of customers in the system is between $\hat{n}_L$ and $\hat{n}_H$; and it does not accept customers of either type if the total number of customers in the system is greater than $\hat{n}_H$. In terms of the firm’s optimal scheduling policy, when we have $h_L = h_H$, we focus on the case where the firm serves the customers in a first-come, first-served manner, regardless of their types.

The above lemmas imply that if the firm has full control over customers’ admission to the system and has full information about the customer types, it is optimal for the firm to adopt the threshold-based policy characterized by the two switching curves $S_H(n^0_L)$ and $S_L(n^0_H)$. Moreover, when the per unit holding cost is the same for all customers, we can simplify these switching curves and characterize the firm’s optimal admission policy by the two finite thresholds $\hat{n}_L$ and $\hat{n}_H$.

4.2. Influential Cheap Talk: homogeneous holding cost

We next construct the equilibria for our model based on the results for the full information and full control case. Note that, in our model, customers have no information about the system status; while the firm not only has no control over customer behavior, but also lacks the ability to differentiate customers of different types. The key questions now are whether and how the firm can credibly communicate with the customers using delay announcements in our model. Given that the firm’s best solution is different when the per unit holding cost is the same for all customers and when per unit holding cost is different for customers of different types, we consider these two cases separately.

In this section, we focus on the case with $h_H = h_L$. We will investigate the case with $h_H \neq h_L$ in Section 4.3.
Note that the firm obtains a higher value by serving a high type customer than by serving a low type customer. Thus, in the case with \( h_H = h_L \), the firm would prefer admitting a high type customer to a low type customer. When the high type customers are more patient than the low type customers, i.e., \( \frac{R_H}{c_H} > \frac{R_L}{c_L} \), high type customers will join the system whenever the low type join. Due to such incentive alignment between the customers and the firm, we show that the firm may be able to achieve the first best solution through announcements when high type customers are more patient than the low type. In order to characterize such an equilibrium, we let \( \bar{n}_L \) be the expected number of customers in the system conditional on the number of customers in the system being less than \( \hat{n}_f^L \) under the first best solution. Similarly, we define \( \bar{n}_H \) as the expected number of customers in the system conditional on the number of customers in the system being between \( \hat{n}_f^L \) and \( \hat{n}_f^H \) under the first best solution. We now construct the equilibrium where the firm achieves the first best solution formally in the following proposition.

**Proposition 1.** When \( h_H = h_L \) and \( \hat{n}_f^H > \hat{n}_f^L \), there exists an equilibrium with influential cheap talk, in which the firm achieves its first best solution, if and only if,

\[
\bar{n}_L + 1 \leq \frac{R_L h}{c_L} < \bar{n}_L + 1, \tag{5}
\]

\[
\bar{n}_H + 1 \leq \frac{R_H h}{c_H} < \bar{n}_H + 2. \tag{6}
\]

Furthermore, one such equilibrium is defined as follows: the announcement policy of the firm is given by

\[
A(n) = \begin{cases} 
  m_1 & \text{if } n \leq \hat{n}_f^L \\
  m_2 & \text{if } \hat{n}_f^L < n \leq \hat{n}_f^H \\
  m_3 & \text{otherwise,}
\end{cases}
\]

customers are served in a first-come, first served manner, and the action rules of low type and high type customers are given by

\[
a_L(m) = \begin{cases} 
  \text{join} & \text{if } m = m_1 \\
  \text{balk} & \text{otherwise,}
\end{cases} \quad a_H(m) = \begin{cases} 
  \text{join} & \text{if } m = m_1 \text{ or } m = m_2 \\
  \text{balk} & \text{otherwise.}
\end{cases}
\]

where \( \hat{n}_f^L \) and \( \hat{n}_f^H \) are the thresholds identified in Lemma 2.

The equilibrium above shows that the firm may be able to achieve the first best solution without fully separating the customers. In particular, the firm uses three announcements to signal three different levels of congestion, i.e., Low, Medium, and High. When the congestion level is low, all customers join the system. When the congestion level is medium, only the high type customers join but not the low type. Meanwhile, when the congestion level is high, neither type of the customers join the system. The solution above is clearly incentive compatible to the firm, as it allows the firm to achieve its first best solution. From the customers’ point of view, as long as their reward-cost
ratios are between the four thresholds given in Proposition 1, customers have no incentive to deviate from the first best solution either. In particular, low type customers obtain positive expected utility when they receive the message \( m_1 \), while obtain negative expected utility otherwise. Similarly, high type customers obtain positive expected utility when they receive the messages \( m_1 \) and \( m_2 \), while obtain negative expected utility otherwise. While this is an influential equilibrium, it is also a semi-separating equilibriums. This is because one of the messages, i.e., \( m_2 \) triggers different reactions from customers of different types, while the messages \( m_1 \) and \( m_3 \) trigger the same reactions from both customer types.

Note that for the firm to achieve the first best solution, it requires the high type customers to be more patient than the low type, i.e., \( \frac{R_H}{c_H} > \frac{R_L}{c_L} \). The next question is whether the firm can replicate the first best solution when the low type customers are more patient than the high type, i.e., \( \frac{R_L}{c_L} > \frac{R_H}{c_H} \). In this case, low type customers are willing to join the system whenever the high type customers are. However, in the case of full information and full control scenario, the firm is willing to admit high type customers whenever it admits low type customers. Due to this opposite preferences of the firm and the customers, the firm cannot achieve the first best solution through delay announcements. In fact, the best the firm can do is to induce an influential pooling equilibrium, where customers of both types react to announcements identically.

According to Definition 3, in a pooling equilibrium, the firm treats customers of different types identically, and customers of different types respond to the announcements in the same manner. Hence, similar to Allon et al. (2011), we can construct the pooling equilibrium as if there is only one type of customers, by using one single threshold, referred to as \( \hat{n} \). We denote the expected number of customers in the system conditional on the number of customers in the system being not larger than \( \hat{n} \) under the pooling equilibrium as \( \bar{n} \). Such an equilibrium is characterized in the following proposition. We show that given \( \frac{R_L}{c_L} > \frac{R_H}{c_H} \), there are no other equilibria, where the firm can achieve a higher profit.

**Proposition 2.** When \( h_H = h_L \), the firm may achieve a pooling equilibrium, if and only if,

\[
\hat{n} + 2 > \frac{R_L \mu}{c_L} > \frac{R_H \mu}{c_H} \geq \bar{n} + 1
\]

One such equilibrium is defined as follows: the announcement policy of the firm is given by

\[
A(n) = \begin{cases} 
m_1 & \text{if } n \leq \hat{n} \\
m_2 & \text{otherwise}
\end{cases}
\]

and the action rules of the customers are given by

\[^4\text{As we will discuss in Section 4.3, Proposition 2 continues to hold when we have } h_H < h_L.\]
\[ a_L(m) = \begin{cases} \text{join} & \text{if } m = m_1 \\ \text{balk} & \text{otherwise} \end{cases}, \quad a_H(m) = \begin{cases} \text{join} & \text{if } m = m_1 \\ \text{balk} & \text{otherwise} \end{cases}. \]

As for the firm’s scheduling policy \( g \), the firm serves customers in a first-come, first-served manner. Furthermore, the firm’s profits under any other equilibria are bounded by the profit under the above pooling equilibrium.

As one may expect, given that there are two types of customers, the firm may want to elicit information from customers regarding their types at least to certain extent in order to maximize profits. However, we show that, the pooling equilibrium, where the firm does not elicit information on customer types at all, may perform the best in the firm’s profit among all other equilibria.

4.3. Influential Cheap Talk: heterogeneous holding cost

We have focused on the scenario where the per unit holding cost is the same for both customer types in Section 4.2. We now turn to the case when the holding cost is different for customers of different types. Recall that the order of service does not impact the firm’s profit when we have \( h_H = h_L \). To this end, the firm focuses on the problem of what announcement to provide to induce the desired customer responses. However, when we have \( h_H \neq h_L \), besides providing delay announcements to influence customers’ decision on whether to join the system, the firm may also like to prioritize the customers who have joined the system appropriately based on their types to reduce its overall cost.

Recall that we have shown that the firm can achieve the first best solution through delay announcements without observing customer types or fully separating the customers when the per unit holding cost is homogeneous among customers in Section 4.2. However, we now claim that the firm cannot achieve its first best solution via delay announcements when the per unit holding costs for customers of different types are different. Note that the firm can only prioritize the customers, whose types it knows. Meanwhile, the firm can only elicit information on customer types, when customers of different types respond to announcements differently. We next argue that the firm cannot fully separate customers of different types through delay announcements. Based on Lemma 1, one can see that, to achieve the first best solution, the firm would like to admit both customer types when there are no customers in the system for any non-degenerate case with \( S_i(0) \geq 0, \forall i \in \{H, L\} \).

As a result, to achieve the first best, the firm must provide at least one message which induces both customer types to join the system. This prevents the firm from fully separating the customers and thus to achieve the first best. We next present this result formally in the following theorem.

**Theorem 2.** When \( h_L \neq h_H \) and the firm does not observe customer types, the firm cannot achieve the first best solution by only using delay announcements.
The questions now are whether and how the firm can credibly communicate with customers through delay announcements when we have \( h_H \neq h_L \). While the firm cannot fully separate the customers or achieve the first best, it may be able to partially separate customers in equilibria. As a result, the firm can prioritize the customers, whose types it elicits through their different reactions towards announcements, to optimize the profit. To characterize the game played and its equilibrium that emerges between the firm and the customers, it is important to note that there are two different scenarios based on which customer type has a higher per unit holding cost. In this section, we focus on the case when the per unit holding cost of the low type customers is higher than that of the high type customers, i.e., \( h_H < h_L \). (We have conducted a similar analysis for the case when we have \( h_H > h_L \) in Appendix A.

Based on Theorem 1, we can focus on the equilibria where the firm uses at most three announcements. Recall that, in this paper, we only focus on the non-degenerate cases where it is optimal for the firm to admit both customer types and its optimal for both customer types to join when there are no customers in the system. To this end, in any equilibria with influential cheap talk, there exits at least one announcement \( m_2 \in M_{HL} \) which induces both customer types to join the system when there are no customers in the system. Meanwhile, one should see that, in any equilibria, the firm would like to provide a message with \( m_0 \in M_{L} \) to induce both customer types to balk when the system is really congested. It is important to note that the firm obtains a higher value by serving the high type customers than by serving the low type customers. Together with the fact that the per unit holding cost of the high type customers is lower than that of the low type customers, the firm may like to provide a message with \( m_1 \in M_{H} \) to induce high type customers to join but low type customers to balk, if such customer response can be sustained in an equilibrium. One should see that this customer response may only be sustained in an equilibrium if the high type customers are more patient than the low type customers. In fact, the following proposition shows that, under certain incentive compatibility conditions on customers’ patience time, there exists a semi-separating equilibrium where the firm induces the following customer responses by providing the corresponding announcements: 1) when the firm provides message \( m_0 \in M_{L} \), neither the low nor the high type customers join the system; 2) when the firm provides announcement \( m_1 \in M_{H} \), only high type customers but not the low type join the system; and, 3) both types of customers join the system when they are provided with message \( m_2 \in M_{HL} \). Note that the firm’s optimal policy is not only comprised of the announcement policy but also the priority policy. In particular, we show that, under the semi-separating equilibrium, it is optimal for the firm to prioritize the customers receiving the message \( m_2 \) over customers receiving the announcement \( m_1 \).

Above, we described the strategy of the firm in terms of both the announcement policy and priority policy under the semi-separating equilibrium. To characterize the corresponding customer
incentive compatibility conditions, we let $\bar{w}_{m_0}$, $\bar{w}_{m_1}$, and $\bar{w}_{m_2}$ denote the expected waiting time of customers receiving the message $m_0$, $m_1$ and $m_2$, respectively, under the semi-separating equilibrium. We next formally present the semi-separating equilibrium in the following proposition.

**Proposition 3.** When $h_H < h_L$, there exists a semi-separating equilibrium with influential cheap talk, if and only if,

$$\bar{w}_{m_2} \leq \frac{R_L}{c_L} < \bar{w}_{m_1} \leq \frac{R_H}{c_H} < \bar{w}_{m_0} \quad (9)$$

Furthermore, one such equilibrium is defined as follows: the action rules of the low and high type customers are given by

$$a_L(m) = \begin{cases} 
\text{join} & \text{if } m = m_2 \\
\text{balk} & \text{otherwise},
\end{cases} \quad a_H(m) = \begin{cases} 
\text{join} & \text{if } m = m_1 \text{ or } m = m_2 \\
\text{balk} & \text{otherwise}.
\end{cases}$$

In terms of the firm’s strategy, the firm provides three distinct messages $m_0$, $m_1$ and $m_2$ which satisfy the condition given by (9). However, we cannot explicitly characterize the announcement policy. The optimal scheduling rule of the firm is given by

$$g(n_H, n_L, n_{HL}) = \begin{cases} 
m_2 & \text{if } n_{HL} > 0 \\
m_1 & \text{if } n_{HL} = 0 \text{ and } n_H > 0 \\
m_0 & \text{if } n_{HL} = n_H = 0
\end{cases}$$

with $n_L = 0$.

It is important to note that the equilibrium above requires the high type customers to be more patient than the low type customers, i.e., $\frac{R_H}{c_H} > \frac{R_L}{c_L}$. The question now is what if we have the low type customers to be more patient than the high type customers. Following a similar argument on the misalignment between the firm and the customers’ preferences in Section 4.2, one can show that the firm achieves the maximum profit in a pooling equilibrium among all equilibria. Under the pooling equilibrium, customers of different types respond to announcements identically. Hence, the firm is not able to elicit information about customers’ types at all in a pooling equilibrium and in turn the firm cannot prioritize customers. As a result, the equilibrium that emerges between the customers and the firm under this scenario is identical to the one characterized in Proposition 2.

### 4.4 Babbling Equilibria

We have focused on the influential equilibrium where the firm provides credible information and customers take the announcements into account when they make abandonment decision. However, in practice, there are many service providers that share no information whatsoever with the customers or information uncorrelated with the state of the system. To this end, we explore whether these systems are in equilibrium. We show that such an equilibrium where no meaningful information is provided by the firm and customers disregard the announcements may indeed exist. In line with the cheap talk literature, we refer to it as a babbling equilibrium, which is formally defined as follows.
Definition 5 (Babbling Equilibrium). We claim that a pure strategy MPBNE \((a_L, a_H, A, g)\) is a babbling equilibrium if the two random variables, i.e., the announcement given by the firm \(A(Q(a_L, a_H, A, g))\) and the system state \(Q(a_L, a_H, A, g)\), are independent, and \(a_i(m_1) = a_i(m_2)\) for all \(i \in \{H, L\}\), \(m_1, m_2 \in M\).

Note that there are two different actions, i.e., join or balk, for a customer from either class. As a result, one may expect that there exits four types of pure strategy babbling equilibria. However, one can show that it cannot be an equilibrium when customers of both types balk. Thus, there are only three types of pure strategy babbling equilibria that may exist: 1) a pure strategy babbling equilibrium where both low and high type customers join the system regardless of the announcements; 2) a pure strategy babbling equilibrium where only high type customers join the system, while all the low type customers balk; 3) a pure strategy babbling equilibrium where only low type customers join the system, while all high type customers balk.

The question now is under what conditions these babbling equilibria may exist. To address this question, we start by exploring the conditions under which the babbling equilibrium where both types of the customers join the system regardless of the announcements may exist. If customers of both types indeed join the queue disregard of the announcements received, the system becomes an \(M/M/1\) system with the arrival rate and the service rate being \(\lambda\) and \(\mu\), respectively. Thus, one can show that the average waiting time in the system is given by \(\frac{1}{\mu - \lambda}\). Since customers would join the system if and only if their expected utility is positive in equilibrium, we have \(R_i - \frac{c_i}{\mu - \lambda} \geq 0 \ \forall i \in \{H, L\}\). Given that the firm cannot differentiate customer types in any way through a babbling equilibrium, we focus on the case when the firm serves the customers in a first-come, first-served manner. Following a similar logic, we can characterize the other two types of pure strategy babbling equilibria. We formalize the characterization in the following proposition.

Proposition 4. 1. The pure strategy babbling equilibrium where both low and high type customers join the system exists, if and only if, \(\frac{R_i}{c_i} \geq \frac{1}{\mu - \lambda}, \forall i \in \{H, L\}\).

2. The pure strategy babbling equilibrium where all high type customers join the system but none of the low type customers do exists, if and only if, \(\frac{R_L}{c_L} < \frac{1}{\mu - \beta H \lambda} \leq \frac{R_H}{c_H}\).

3. The pure strategy babbling equilibrium where all low type customers join the system but none of the high type customers do exists, if and only if, \(\frac{R_H}{c_H} < \frac{1}{\mu - \beta L \lambda} \leq \frac{R_L}{c_L}\).

Based on the proposition above, one can see that none of these pure strategy babbling equilibria can co-exist. Moreover, neither the firm’s value of serving customers nor its holding cost impacts the existence of any of the babbling equilibria.
4.5. Should the firm provide announcements?

We have shown that both the babbling equilibria and the influential equilibria may exist. The
question now is which equilibrium the firm and the customers would prefer. To this end, we compare
the influential equilibria with the babbling ones in the regions where they both exist, in terms
of both customers’ utility and the firm’s profit. Note that there exists two types of influential
equilibria, i.e., the semi-separating equilibrium and the pooling equilibrium. Meanwhile, we have
three types of babbling equilibrium characterized in Proposition 4. Let us start with the comparison
between the pooling equilibrium and the babbling equilibrium. Given that the babbling equilibria
are mutually exclusive, there is at most one babbling equilibrium which may co-exist with the
pooling equilibrium for given parameters. To this end, we let $\Pi_{IP}$ and $U_{oIP}$ denote the profit of the
firm and the overall total customers’ utility in the pooling equilibrium, respectively. Moreover, we
refer to $\Pi_{NI}$ and $U_{oNI}$ as the profit of the firm and the overall total customers’ utility in the babbling
equilibrium which co-exists with the pooling equilibrium for the given parameters. The following
proposition shows that the firm achieves a higher profit under the pooling equilibrium compared to
the one achieved in the corresponding babbling equilibrium. Moreover, from customers’ perspective,
customers obtain a higher overall total utility in the pooling equilibrium compared to the one
obtained in the babbling equilibrium. We now present the above results rigorously in the following
proposition.

**Proposition 5.** Assume that both a pure strategy pooling equilibrium with influential cheap talk
and a pure strategy babbling equilibrium exist, then we have:

1. $\Pi_{NI} < \Pi_{IP}$;
2. $U_{oNI} < U_{oIP}$;

We obtain similar results when we compare the semi-separating equilibrium and the babbling
equilibria through extensive numerical studies. These results imply that providing delay announce-
ments not only improves firm’s profit but also the overall customers’ utility compared to the case
when announcements are not provided. Note that, under the pooling equilibrium, the firm’s profit
and the overall customer utility in our system with two customer classes are the same as the ones
in a system with one single customer class, whose reward of service and per unit waiting cost are
$\beta_H R_H + \beta_L R_L$ and $\beta_H C_H + \beta_L C_L$, respectively. To this end, the intuition for the results above is
similar to the one presented in Allon et al. (2011). In particular, Naor (1969) shows that customers
are more willing to join the system than what the social planner would like them to. This is because
customers make decision on whether to join only to maximize their own utility, while ignore the
negative externalities that they may impose on other customers by joining the system. The thresh-
old that the firm induces through the pooling equilibrium helps reduce such externalities and thus
improves the overall customer utility.
While providing delay announcements improves the overall customer utility, we next show that it may improve or hurt the utility of the more patient customer type. We illustrate this result via two numerical examples.

**Example 1:** In this example, we let the high type customers be more patient than the low type customers, i.e., \( R_H > c_H > R_L > c_L \). In particular, we assume the reward of receiving service for both customer types to be unity, i.e., \( R_H = R_L = 1 \), while the per minute waiting cost of the high type customers and that of the low type customers are assumed to be 0.2 and 2.9, respectively, i.e., \( c_H = 0.2 \) and \( c_L = 2.9 \). Furthermore, we let the total arrival rate \( \lambda \) be 6.7 customers per minute while the service rate \( \mu \) be 8 customers per minute. The value of serving a high and a low type customer for the firm are assumed to be 100 and 2, respectively, while the per unit holding cost is assumed to be one, i.e., \( v_H = 100 \), \( v_L = 2 \) and \( h = 1 \). Lastly, the percentage of the high type customers is assumed to be 90\%, i.e., \( \beta_H = 90\% \). We evaluate the firm’s optimal policy using value iteration over a truncated state space. Given this optimal policy, one can show that, if the firm provides announcements to its customers, the firm can achieve the first best through a semi-separating equilibrium characterized in Proposition 1. Otherwise, there exists a babbling equilibrium where only high type customers join the system, which is characterized in Proposition 4. We then evaluate the utility per unit time for the high type customers under both the semi-separating equilibrium and the babbling equilibrium. We show that the utilities per hour of the high type customers under the semi-separating equilibrium and the babbling equilibrium are equal to 331 and 325, respectively. Thus, in this example, we show that utility of the more patient customer type (high type) under the influential equilibrium is higher than the one under the babbling equilibrium.

One might expect the utility of the more patient customer type to be higher when delay announcements are provided as shown in the example above. However, it is important to note that more of the less patient customers would join the system when announcements are provided compared to the case when announcements are not provided. To this end, providing announcements may impose more negative externalities on the more patient customers due to the increased number of the less patient customers in the system. As a result, providing delay announcements may hurt the utility of the more patient customer type compared to the case when announcements are not provided, see **Example 2**.

**Example 2:** In this example, we use the same parameters as in **Example 1** with the following modification: \( \beta_H = 70\% \) and \( \mu = 7.5 \) per minute. Similar to **Example 1**, one can show that, if the firm provides announcements to its customers, the firm can achieve the first best through a semi-separating equilibrium characterized in Proposition 1. Otherwise, there exists a babbling equilibrium where only high type customers join the system, which is characterized in Proposition
4. We then evaluate the utility per unit time for the high type customers under both the semi-separating equilibrium and the babbling equilibrium. We show that the utilities of the high type customers per hour under the semi-separating equilibrium and the babbling equilibrium are equal to 260 and 262, respectively. Thus, in this example, we show that utility of the more patient customer type (high type) under the influential equilibrium is lower than the one under the babbling equilibrium.

To sum up, we show that providing delay announcements increases the firm’s profits. Meanwhile, from the customer perspective, providing delay announcements improves the overall customer utility, but it may improve or hurt the utility of the more patient customer type. These results are consistent with the empirical results shown in Yu et al. (2015). Note that Yu et al. (2015) focuses on the scenario where the firm’s announcement policy is fixed, while the firm is strategic in its announcement policy to maximize profits in this paper. This implies that the results above are not driven by the cheap talk but rather by the role the announcements play in encouraging customers to join or not.

5. Model with Information on Customer Types

So far, we assume that the firm cannot observe customer types. However, it is important to note that, in practice, many firms have information on customer types, (e.g., call centers request customers to reveal service types, online retailers know customer types through their registered accounts, etc.) To study the value that the firm may gain or lose when the firm has the information about customer types upon their arrivals, we now extend our model by allowing the firm to observe customer types before it provides announcements. We refer to this model as the model with information. This model is identical to the model with no information presented in Section 3 with two key modifications: 1) the firm can now decide on whether to provide announcements and what announcements to provide to customers based on their types; and 2) the firm can schedule customers based on their types instead of the announcements that they receive. To incorporate these changes in the model with information, we first let $S_I$ represent the set of system states in the model with information. Given the firm has perfect information on customer types in this model, the system states can be characterized by the number of low type customers $n^0_L$ and the number of high type customers $n^0_H$. Thus, we have the set of the system states $S_I$ given by $S_I = \{(n^0_H, n^0_L) | (n^0_H, n^0_L) \in \mathbb{Z}^2\}$, which coincides with the set of system states for the full information and full control case presented in Section 4.1. We then let the announcement policy of the firm to type $i$ customers be a function given by $A^i : S_I \mapsto M$ with $i \in \{H, L\}$. To account for the new feature on the firm’s scheduling policy, we let the scheduling policy of the firm be given by a function $g_i : S_I \mapsto \{\emptyset, L, H\}$. In particular, we have $g_i(n^0_H, n^0_L) = i \in \{H, L\}$, if the next customer
to be served is the first customer of type $i$ in the system, when there are $n_H^0$ high type customers and $n_L^0$ low type customers in the system. Meanwhile, we have $g_I(n_H^0, n_L^0) = \emptyset$, if the firm decides to be idle when there are $n_H^0$ high type customers and $n_L^0$ low type customers in the system. It is worth mentioning that the subscription $I$ in $S_I$ and $g_I$ indicates the condition that the firm has information on customer types. Recall that, when the firm does not observe the types of the customers, the set of system states is given by $S = \{(n_H, n_L, n_{HL}) | (n_H, n_L, n_{HL}) \in \mathbb{Z}^3\}$, where $n_H$, $n_L$, and $n_{HL}$ are the number of customers in the system receiving announcements from the subsets $M_{(H)}$, $M_{(L)}$ and $M_{(HL)}$, respectively. As for the scheduling policy of the firm in the model with no information, the firm can only schedule the customers based on the announcements that they receive. In particular, the scheduling policy is given by the function $g : S \mapsto X$, where $S$ is the set of system states and $X \in \{M_\emptyset, M_{(H)}, M_{(L)}, M_{(HL)}\}$ is the type of announcements that customers receive.

To characterize the system dynamics, we let $p_l(n_H^0, n_L^0 | a_H, a_L, g_I, A^H, A^L)$ be the probability that there are $n_H^0$ high type and $n_L^0$ low type customers in the system in the steady state given the customers’ strategy $a_i$, the firm’s scheduling rule $g_I$ and announcement policy $A^i$ with $i \in \{H, L\}$. To define the equilibria that emerge between the firm and the customers in the model with information, we employ the equilibrium concept of MPBNE which is the same as the one used in the model with no information. Meanwhile, we let $w_H^0(n_H^0, n_L^0)$ be the waiting time of the high type customer who joins the system when there are $n_H^0$ high type and $n_L^0$ low type customers in the system under the equilibrium. Similarly, $w_L^0(n_H^0, n_L^0)$ is the waiting time of the low type customer who joins the system when the system state is $(n_H^0, n_L^0)$. We next formally define the pure strategy equilibrium for the model with information in the following definition.

**Definition 6.** We say that $(a_H, a_L, g_I, A^H, A^L)$ forms a Markov Perfect Bayesian Nash Equilibrium (MPBNE), if and only if, it satisfies the following conditions:

1. For each $m \in M$ and $i \in \{L, H\}$, we have

   $$a_i(m) = \begin{cases} 1 & \text{if } \frac{\sum_{(n_H^0, n_L^0) | A^i(n_H^0, n_L^0) = m} [R_I - c_i w_H^0(n_H^0, n_L^0)] p(n_H^0, n_L^0 | a_H, a_L, g_I, A^H, A^L)}{\sum_{(n_H^0, n_L^0) | A^i(n_H^0, n_L^0) = m} p(n_H^0, n_L^0 | a_H, a_L, g_I, A^H, A^L)} \geq 0 \end{cases} \tag{10}$$

2. There exists value functions $V_I(n_H^0, n_L^0)$ with $(n_H^0, n_L^0) \in \mathbb{Z}^2$, constant $\gamma_I$, and the announcement policy $m^i = A^i(n_H^0, n_L^0)$ that solve the following equation:

   $$V_I(n_H^0, n_L^0) + \frac{\gamma_I}{\lambda} = \frac{1}{\lambda} \left\{ -h_L n_L^0 - h_H n_H^0 + \beta_H \lambda \max_{m_H \in M} \left\{ V_I(n_H^0, n_L^0)(1 - a_H(m)) + (V_I(n_H^0 + 1, n_L^0) + v_H) a_H(m) \right\} \right\}$$
\[ + \beta_L \lambda \max_{m \in M} \left\{ V_i(n_H^0, n_L^0)(1 - a_L(m)) + (V_i(n_H^0, n_L^0 + 1) + v_L)a_L(m) \right\} \]

\[ + \mu \max \left\{ V_i(n_H^0 - 1, n_L^0)I_{\{n_H^0 > 0\}} + V_i(n_H^0, n_L^0)I_{\{n_H^0 = 0\}} \right\}, \]

\[ V_i(n_H^0, n_L^0 - 1)I_{\{n_H^0 > 0\}} + V_i(n_H^0, n_L^0)I_{\{n_H^0 = 0\}}, \]

\[ V_i(n_H^0, n_L^0) \right\}, \]

(11)

with \( \Lambda = \lambda + \mu \). Thus, the firm’s optimal announcement policy is given by \( m^i = A_i^0(n_H^0, n_L^0) \) with \( i \in \{H, L\} \).

3. The firm’s optimal scheduling policy is given by

\[
g_i(n_H^0, n_L^0) = \begin{cases} H & \text{if } V_{i1} = \max \{V_{i1}, V_{i2}, V_{i3}\} \\ L & \text{if } V_{i2} = \max \{V_{i1}, V_{i2}, V_{i3}\} \\ \emptyset & \text{if } V_{i3} = \max \{V_{i1}, V_{i2}, V_{i3}\}, \end{cases} \]

(12)

where

\[ V_{i1} = V_i(n_H^0 - 1, n_L^0)I_{\{n_H^0 > 0\}} + V_i(n_H^0, n_L^0)I_{\{n_H^0 = 0\}}, \]

\[ V_{i2} = V_i(n_H^0, n_L^0 - 1)I_{\{n_H^0 > 0\}} + V_i(n_H^0, n_L^0)I_{\{n_H^0 = 0\}}, \]

and

\[ V_{i3} = V_i(n_H^0, n_L^0). \]

The above definition is related to the one defined for the model with no information, see Definition 1. The key difference is that, in the model with information, the firm can provide announcements and schedule customers based on the type of the customers. These unique features in the firm’s announcement policy and scheduling policy are captured in (11) and (12), respectively.

### 5.1. Equilibria with Information

We next explore the equilibria that emerge between the customers and the firm when the firm observes customer types upon their arrivals. Note that we have characterized the first best solution of the firm in Section 4.1, where the firm has full information about customer types and full control over customer admission. Although in our model, the firm does not have control over customers’ admission, the following proposition shows that the queuing dynamic observed under any MPBNE with influential cheap talk (if it exists) corresponds to the one where the firm achieves its first best solution. Note that, we say an MPBNE is influential if the announcements are influential for both customer types in the model with information. This is in line with the definition on influential cheap talk for the model with no information, see Definition 2.

**Theorem 3.** When the firm observes customer types, the firm achieves its first best solution under any MPBNE with influential cheap talk.
Based on the theorem above, to construct the equilibrium when the firm observes customer types, we consider the system where the firm implements the first best solution. Note that we have characterized the first best solution, where the firm has full information on customer types and full control over customer admission in Lemmas 1 and 2.

Recall that, to achieve the first best, the firm would like the high type customers to join the system when the number of high type customers in the system is not larger than the threshold \( S_H(n_H^0) \), i.e., \( n_H^0 \leq S_H(n_H^0) \). Otherwise, the firm would like the high type customers to balk. Similarly, the firm would like to accept the low type customers when the number of the low type customers is not larger than the threshold \( S_L(n_L^0) \), i.e., \( n_L^0 \leq S_L(n_L^0) \). Otherwise, the firm would like the low type customers to balk. To characterize the equilibrium, we let \( \bar{w}_i \) and \( \bar{w}_i \), with \( i \in \{H,L\} \), be the expected waiting time of the arriving type \( i \) customer (if she joins the system) given that the firm wants her to join and balk the system, respectively. Note that \( w_H^0(n_H^0,n_L^0) \) denotes the waiting time of the high type customer who joins the system when there are \( n_H^0 \) high type and \( n_L^0 \) low type customers in the system. Similarly, \( w_L^0(n_H^0,n_L^0) \) denotes the waiting time of the low type customer who joins the system when the system state is \( (n_H^0,n_L^0) \). To this end, we have

\[
\bar{w}_H = \frac{\sum_{n_L^0=0}^{\infty} \sum_{n_H^0=0}^{S_H(n_H^0)} w_H^0(n_H^0,n_L^0)p_I(n_H^0,n_L^0|a_H,a_L,g_I,A^H,A^L)}{\sum_{n_L^0=0}^{\infty} \sum_{n_H^0=0}^{\infty} p_I(n_H^0,n_L^0|a_H,a_L,g_I,A^H,A^L)},
\]

\[
\bar{w}_L = \frac{\sum_{n_L^0=0}^{\infty} \sum_{n_H^0=0}^{\infty} S_H(n_H^0) w_H^0(n_H^0,n_L^0)p_I(n_H^0,n_L^0|a_H,a_L,g_I,A^H,A^L)}{\sum_{n_L^0=0}^{\infty} \sum_{n_H^0=0}^{\infty} S_H(n_H^0) p_I(n_H^0,n_L^0|a_H,a_L,g_I,A^H,A^L)},
\]

\[
w_L = \frac{\sum_{n_L^0=0}^{\infty} \sum_{n_H^0=0}^{\infty} S_L(n_L^0) w_L^0(n_H^0,n_L^0)p_I(n_H^0,n_L^0|a_H,a_L,g_I,A^H,A^L)}{\sum_{n_L^0=0}^{\infty} \sum_{n_H^0=0}^{\infty} S_L(n_L^0) p_I(n_H^0,n_L^0|a_H,a_L,g_I,A^H,A^L)},
\]

\[
w_L = \frac{\sum_{n_L^0=0}^{\infty} \sum_{n_H^0=0}^{\infty} S_L(n_L^0) w_L^0(n_H^0,n_L^0)p_I(n_H^0,n_L^0|a_H,a_L,g_I,A^H,A^L)}{\sum_{n_L^0=0}^{\infty} \sum_{n_H^0=0}^{\infty} S_L(n_L^0) p_I(n_H^0,n_L^0|a_H,a_L,g_I,A^H,A^L)},
\]

Note that we have \( \bar{w}_H = \frac{\hat{h}_H+2}{m} \) and \( \bar{w}_L = \frac{\hat{n}_L+1}{m} \) for the case with \( h_H = h_L \), where \( \hat{n}_L \) and \( \hat{n}_H \) are the thresholds given in (5) and (6), respectively. We next characterize the equilibrium where the firm achieves the first best while observing customer types upon their arrivals in the following proposition.

**Proposition 6.** There exists an equilibrium with influential cheap talk where the firm achieves the first best, if and only if,

\[
w_i \leq \frac{R_i}{c_i} \leq \bar{w}_i, \quad \forall i \in \{H,L\}
\]

(13)
Furthermore, one such equilibrium is defined as follows: The announcement policy of the firm is given by
\[
A^H(n^0_H, n^0_L) = \begin{cases} 
\frac{m^H_1}{m^H_2} & \text{if } n^0_H \leq S^H_H(n^0_L) \\
\frac{m^L_1}{m^L_2} & \text{otherwise}
\end{cases}
\]
\[
A^L(n^0_H, n^0_L) = \begin{cases} 
\frac{m^L_1}{m^L_2} & \text{if } n^0_L \leq S^L_L(n^0_H) \\
\frac{m^L_1}{m^L_2} & \text{otherwise}.
\end{cases}
\]

Moreover, the action rules of low and high type customers are given by
\[
a^H_H(m) = \begin{cases} 
\text{join} & \text{if } m = m^H_1 \\
\text{balk} & \text{if } m = m^H_2
\end{cases}
\]
\[
a^L_L(m) = \begin{cases} 
\text{join} & \text{if } m = m^L_1 \\
\text{balk} & \text{if } m = m^L_2
\end{cases}
\]

As for the scheduling policy of the firm, it serves customers with the same per unit holding cost in a first-come, first-served manner. When \(h_H \neq h_L\), the firm’s optimal scheduling policy is given as follows:
\[
g^I(n^0_H, n^0_L) = \begin{cases} 
K_1 & \text{if } n^0_{K_1} > 0 \\
K_2 & \text{if } n^0_{K_1} = 0 \text{ and } n^0_{K_2} > 0 \\
\emptyset & \text{if } n^0_{K_1} = n^0_{K_2} = 0,
\end{cases}
\]
with \(K_1 = \arg\max_{i \in \{H, L\}} h_i\) and \(K_2 = \arg\min_{i \in \{H, L\}} h_i\).

Note that the firm clearly has no incentive to deviate from the first best solution. As for the customers, due to incentive compatible conditions given in (13), it is optimal for them to follow the first best solution prescribed by the firm.

5.2. Comparison: Information vs No Information

In this section, we contrast the equilibria that emerge when the firm can observe customer types to the ones when the firm does not. In particular, we explore whether the firm can improve its capability to influence customers by observing customer types upon their arrivals and if so, under what conditions. Note that the equilibria that emerge in the cases with \(h_H = h_L\), \(h_H < h_L\) or \(h_H > h_L\) may be all different. Thus, we shall compare the equilibria in the model where the firm does not observe customer types to the ones in the model with information for each of these three cases separately. However, the insights that we obtain from the case with \(h_H > h_L\) through such comparison is similar to the case with \(h_H < h_L\). To this end, we will focus on the cases with \(h_H = h_L\) and \(h_H < h_L\) in this section.

Following Theorem 3 and Proposition 6, the necessary and sufficient condition for the existence of equilibria with influential cheap talk can also be written as \(\frac{R_i}{c_i} \in [\bar{w}_i, \bar{w}_i], \forall i \in \{H, L\}\) in the model with information. We can view \(\frac{R_i}{c_i}\) as the type \(i\) customers’ perspective on their willingness to wait, while \(\bar{w}_i, \bar{w}_i\) as the firm’s perspective on the desired congestion level of the system for type \(i\) customers with \(i \in \{H, L\}\). In studying the impact of the firm observing customer types, we shall fix the firm’s perspective and vary the customers’ perspective. In particular, we introduce the following terminology: for given fixed firm’s cost parameters, the ratio of each customer type, the service and
arrival rate, we let $D_I$ and $D_{NI}$ be the set of the patiences of both customer types for which the firm can achieve equilibria with influential cheap talk with and without information on customer types, respectively. Based on the above discussion, we have $D_I = \{ ( \frac{R_L}{c_L}, \frac{R_H}{c_H}) | \frac{R_L}{c_L} \in [\bar{w}_I, \bar{w}_I], \forall i \in \{H, L\} \}$. Figure 1a shows the region $D_I$ for the case with $h_H = h_L$, where the horizontal and vertical axes represent the patiences of the low and high type customers, respectively. Note that when the firm cannot observe customer types, the firm can achieve equilibria with influential cheap talk through either a semi-separating equilibrium or a pooling equilibrium. To this end, we let $D_{NI}^{SS}$ and $D_{NI}^P$ be the set of patiences of both customer types for which the firm achieves the semi-separating equilibrium and the pooling equilibrium without observing customer types, respectively. To this end, we have $D_{NI} = D_{NI}^{SS} \cup D_{NI}^P$. Following Propositions 1 and 3, we have

$$D_{NI}^{SS} = \left\{ \left( \frac{R_L}{c_L}, \frac{R_H}{c_H} \right) | \bar{w}_m \leq \frac{R_L}{c_L} < \bar{w}_m < \frac{R_H}{c_H} \right\} \text{ if } h_H < h_L$$

$$D_{NI}^P = \left\{ \left( \frac{R_L}{c_L}, \frac{R_H}{c_H} \right) | \bar{w}_m \leq \frac{R_L}{c_L} < \frac{R_H}{c_H} \right\} \text{ if } h_H = h_L$$

Moreover, based on Propositions 2, we have $D_{NI}^{SS} = \{ ( \frac{R_L}{c_L}, \frac{R_H}{c_H} ) | \frac{R_H}{c_H} < \frac{R_L}{c_L} < \frac{h^2}{\mu^2} \}$. Figure 1b shows the regions $D_{NI}^{SS}$ and $D_{NI}^P$, juxtaposed with the region $D_I$ depicted in Figure 1a for the case with $h_H = h_L$.

We next define the expansion region due to the information on customer types as $D_I \cap D_{NI}$, where $D_{NI}$ represents the complement of the set $D_{NI}$. Similarly, we define the contraction region due to customer type information as $D_{NI} \cap D_I$, where $D_I$ is the complement of set $D_I$. Lastly, we define the neutral region due to the information on customer types as $D_I \cap D_{NI}$. We say that information on customer types leads to a contraction if the expansion region is empty. Similarly, we say that information on customer types results in an expansion if the contraction region is empty. Lastly, we say that information on customer types leads to a mixed contraction-expansion if neither of these sets is empty. In fact, Figure 1 depicts a case where customer type information results in a mixed contraction-expansion when we have $h_H = h_L$.

The following proposition shows that information on customer types may lead to an expansion or a mixed contraction-expansion when we have $h_H = h_L$. In particular, the expansion region is never empty when we have $h_H = h_L$, while the contraction region may be empty under certain conditions.

**Proposition 7.** When $h_H = h_L$, we have:

1. $D_{NI}^c \cap D_I \neq \emptyset$.
2. $D_I^c \cap D_{NI} = \emptyset$, if and only if, we have $\mu \bar{w}_H \leq \bar{n} + 1 \leq \bar{n}_I + 2 \leq \mu \bar{w}_L$.

It is intuitive that information on customer types may enhance the credibility of the firm by extending the region where the firm achieves the equilibria with influential cheap talk. This is
because when the firm observes customer types, the firm can provide information to customers based on their types to better match their expectation. However, we also find that there might be a contraction region as well. The key reason is that when the firm observes customer types, it will intend to extract more profits from the customers. This may lead to the misalignment between the incentive of the firm and the customers. As a result, the firm fails to achieve an equilibrium with influential cheap talk when it can observe customer types in the contraction region.

Above we focused on the case with $h_H = h_L$, where we show that information on customer types may lead to an expansion, but never a contraction. However, our results show that, when we have $h_H < h_L$, information on customers types may lead to a contraction, but never an expansion. In particular, the following proposition shows that the contraction region is never empty, while the expansion region may be empty under certain conditions.

**Proposition 8.** Assuming $h_H < h_L$, we have

1. $D_{cNI} \cap D_J \neq \emptyset$, if and only if, we have $\bar{w}_{m_2} \leq w_L \leq \bar{w}_L \leq \bar{w}_{m_1} \leq w_H \leq \bar{w}_H \leq w_{m_0}$.
2. $D_f^c \cap D_{NI} \neq \emptyset$.

**5.2.1. Value of Information** Above we explored the question of whether information on customer types would improve or hurt the credibility of the firm. We next study if the creation of credibility translates into the creation of value for the firm.

Note that the firm achieves the first best solution in the expansion or neutral regions when it can observe customer types. To this end, one can see that information on customer types improves the firm’s profits. Similarly, one may expect information on customer types to hurt the profit of the
firm in the contraction region. However, we find that information on customer types may improve or hurt the profit of the firm. To illustrate this result, we present the following numerical examples.

**Example 3:** In this example, we let the total arrival rate $\lambda$ to be 6.7 customers per unit time. There is a single agent whose service rate is 7.5 customers per unit time, i.e., $\mu = 7.5$. We let the value for the firm by serving a high type customer be 15, while the value by serving a low type customer be 10, i.e., $v_H = 15$ and $v_L = 10$. Meanwhile, the per unit holding cost incurred to the firm for the high and low type customers are 1 and 2, respectively. We assume 50% of the customers are low type customers, i.e., $\beta_L = 50\%$. As for customers’ parameters, we let the service value obtained by each of the high and low type customers be 1.3 and 2.1, i.e., $R_H = 1.3$ and $R_L = 2.1$, respectively. Meanwhile, the per unit time waiting costs for the high and low type customers are assumed to both equal 1. One can show that, given the parameters above, when the firm does not have information on customers types, the firm can achieve the pooling equilibrium characterized in Proposition 2. When the firm has information on customer types, it cannot induce any equilibria with influential cheap talk. However, there may exist a babbling equilibrium where both customer types join the system regardless of the announcements received, while the firm gives absolute priority to the low type customers over high type customers.\(^5\) To this end, one can see that the given patiences of the customers belong to the contraction region. We then evaluate the firm’s profit in both the pooling equilibrium and the babbling equilibrium. Our results show that the firm’s profit under the pooling equilibrium is 75 per unit time, while the firm’s profit under the babbling equilibrium is 81 per unit time. Thus, in this example, we show that information on customer types may even improve the profit of the firm in the contraction region.

**Example 4:** In this example, we use the same parameters as the ones in Example 3 with the following modification: $h_L = 3$, $\beta_L = 90\%$ and $R_H = 0.67$. Similar to Example 3, one can show that, if the firm does not observe customer types, the firm can induce the pooling equilibrium characterized in Proposition 2. The firm’s profit under this equilibrium is 58 per unit time. Meanwhile, when the firm can observe customer types, the firm cannot achieve any equilibria with influential cheap talk. However, there may exist a babbling equilibrium where only low type customers join regardless of the announcements while all high type customers balk. This babbling equilibrium is characterized in Proposition 4. The firm’s profit under this babbling equilibrium is 48 per unit time. Based on the above discussion, we also see that the given customer patiences belong to the contraction region. Thus, this example shows that information on customer types could also hurt the firm’s profit in the contraction region.

\(^5\) When the firm observes customer types, a babbling equilibrium where customers of both types join the system regardless of the announcements exists, if and only if, $\frac{R_L}{v_L} \geq \frac{1}{\mu - \beta_L \lambda}$ and $\frac{R_H}{v_H} \geq \frac{1 - \beta_L}{(\mu - \beta_L \lambda)(\mu - x)}$. 

---

Title: Managing Customer Expectations and Priorities in Service Systems
Given that customer type information diminishes the firm’s ability to induce the desired customer response in the contraction region, one may expect information on customer types to hurt the profit of the firm. However, surprisingly, the above examples demonstrate that loss of credibility may even improve the profit for the firm. This is because information on customer types allows the firm to better prioritize the customers in the babbling equilibria. The improvement in profits from the prioritization may more than compensate for the loss due to firm’s lack of ability to induce the desired customer response in the contraction region. As for the customer utility, information on customer types may improve or hurt the customer utility in the contraction, neutral or expansion region.

6. Conclusion

In this paper, we study how to use delay announcements to manage customer expectations and priorities in the presence of heterogeneous customers. We examine this problem by developing a framework which characterizes the strategic interaction between the self-interested firm and heterogeneous selfish customers. We first explored a model where both the customers and the firm have private information of their own. The customers have private information on their types, while the firm has private information on the system status. To study the value that the firm may gain or lose by observing customer types, we also investigated a model where the firm can observe customer types. We characterize the equilibria that emerge between the firm and its diverse customers in both models.

The analysis of the emerging equilibria demonstrates the role of suppressed information in sustaining an equilibrium with influential cheap talk. Our analysis also underscores that the heterogeneity among the customers raises interesting issues about the ability of the firm to influence the different types of customers differently through delay announcements. We show that the firm cannot fully separating the customers of different types through delay announcements and prioritization. This prevents the firm from achieving the first best solution when the per unit holding costs are different for customers of different types. However, the ability to partially separate among the different customer types through delay announcement allows the firm to sustain a semi-separating equilibrium with influential cheap talk to improve profits. Under such semi-separating equilibrium, we show it is optimal for the firm to give absolute priority to customers receiving announcements corresponding to the smallest expected per unit holding cost over customers receiving announcements associated with larger expected per unit holding cost. It is also worth mentioning that, when the per unit holding cost is the same for customers of both types, the firm can achieve the first best solution without fully separating the customers but by only partially separating the customers. Moreover, we show that it improves the profit and the total overall customer utility by providing
delay announcements, but it may hurt the utility of the more patient customer type. To explore the
value that the firm may gain or lose by observing the type of the customer, we have also studied
a model where the firm can observe the types of customers. We show that the information on cus-
tomer types may enhance the firm’s credibility by extending the region where the firm can achieve
equilibria with influential cheap talk. However, such information may also hurt the credibility of
the firm by contracting the region where the firm achieves the pooling equilibrium. We show that
the creation of credibility in the expansion region improves the firm’s profit. Similarly, one may
expect the loss of credibility in the contraction region to hurt the profit of the firm. However, we
show that the loss of credibility may even improve the firm’s profit.

References
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.
Armony, M., N. Shimkin, W. Whitt. 2009. The impact of delay announcements in many-server queues with
Bertsekas, Dimitri P, Dimitri P Bertsekas, Dimitri P Bertsekas, Dimitri P Bertsekas. 2012. Dynamic pro-
gramming and optimal control, vol. 2. Athena Scientific Belmont, MA.
Chen, Feng, Vidyadhar G Kulkarni. 2007. Individual, class-based, and social optimal admission policies in
Econometric Society 1431–1451.
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
Ibrahim, R., W. Whitt. 2009. Real-time delay estimation in overloaded multiserver queues with abandon-
Iravani, Seyed MR, Tieming Liu, David Simchi-Levi. 2012. Optimal production and admission policies in
make-to-stock/make-to-order manufacturing systems. Production and Operations Management 21(2)
224–235.
Knudsen, Chr N. 1972. Individual and social optimization in a multiserver queue with a general cost-benefit


Rue, Robert C, Matthew Rosenshine. 1981. Optimal control for entry of many classes of customers to an m/m/1 queue. Naval Research Logistics Quarterly 28(3) 489–495.


Appendix A: Model with No Information: heterogeneous holding cost with $h_H > h_L$

Recall that, in Section 4.3, we studied whether and how the firm can credibly communicate with customers through delay announcements when the holding costs are different for customers of different types. In particular, we focused on the case when the per unit holding cost of high type customers is smaller than that of the low type customers, i.e., $h_H < h_L$. In this appendix, we next explore the case when we have $h_H > h_L$.

To characterize the influential equilibria, without loss of generality, we consider the setting where the firm uses at most three different announcements. Before we start constructing the equilibrium, note that, when we have $h_H > h_L$, there exists no $m_H \in M_H$ which induces the high type customers to join and the low type customers to balk in any influential equilibrium. To explore the intuition, similar to the case with $h_H < h_L$, for any given influential equilibrium, there exists an announcement $M_{HL} \in M_{HL}$ which induces both customer types to join the system. If there also exists an announcement $m_H \in M_H$ which induces the high type customers to join but low type customers to balk in an influential equilibrium, one can show that the firm would like to prioritize customers receiving the announcement $m_H$ over the customers receiving the announcement $m_{HL}$ in any influential equilibria. This is because the expected per unit holding cost of customers receiving announcement $m_H$ is larger than that of customers receiving the message $m_{HL}$ when we have $h_H > h_L$. To this end, the expected waiting time of customers receiving announcement $m_H$ is shorter than that of customers receiving message $m_{HL}$. Thus, given it is better off for the low type customers to join the system when they receive the message $m_{HL}$, it should also be better off for them to join the system upon receiving the message $m_H$ in the given influential equilibrium. This cannot be the case by the definition of $m_H$, which explains why the customer response that only high type customers join but low type customers balk cannot be sustained in any influential equilibrium.

Based on the above discussion, one can see that there exists no $m_H \in M_H$ in any influential equilibrium for the case with $h_H > h_L$. Meanwhile, as we mentioned above, there exits at least one announcement $m_{HL} \in$...
$M_{(HL)}$ which induces both customer types to join the system when there are no customers in the system. One can also see that, in any influential equilibrium, the firm would like to provide a message with $m_{\emptyset} \in M_{\emptyset}$ to induce both customer types to balk when the system is really congested. Moreover, when the gain due to a lower holding cost for the low type compared to the high type customers more than compensates the loss due to a lower value of serving the low type customer, the firm may like to provide an announcement $m_{L} \in M_{(L)}$ to induce the low type customers to join but high type customers to balk in an influential equilibrium. Note that such customer response can only be sustained when the low type customers are more patient than the high type customers. In fact, we find that, under certain incentive compatibility conditions on customers’ patience time, there exists a semi-separating equilibrium where the firm provides announcement $m_{HL}$, $m_{L}$ and $m_{\emptyset}$ to induce the corresponding customer response described above. Moreover, we show that under this semi-separating equilibrium, it is optimal for the firm to prioritize the customers who receive the message $m_{HL}$ over customers receiving the announcement $m_{L}$. Note that the expected per unit holding cost of customers receiving the message $m_{HL}$ is higher than that of the customers receiving the message $m_{L}$, assuming $h_{H} > h_{L}$. Thus, prioritizing customers receiving the message $m_{HL}$ over customers receiving the announcement $m_{L}$ minimizes the overall cost.

Above we described the strategy of the firm in terms of both the announcement policy and priority policy under the semi-separating equilibrium. To characterize the corresponding customer incentive compatibility conditions, we let $\bar{w}_{m_{\emptyset}}$, $\bar{w}_{m_{L}}$, and $\bar{w}_{m_{HL}}$ denote the expected waiting time of customers receiving the message $m_{\emptyset}$, $m_{L}$ and $m_{HL}$, respectively, under the semi-separating equilibrium. We next formally present the semi-separating equilibrium in the following proposition.

**Proposition 9.** When $h_{H} > h_{L}$, there exists a semi-separating equilibrium with influential cheap talk, if and only if,

$$\bar{w}_{m_{HL}} \leq \frac{R_{H}}{c_{H}} < \bar{w}_{m_{L}} \leq \frac{R_{L}}{c_{L}} < \bar{w}_{m_{\emptyset}}.$$ 

Furthermore, one such equilibrium is defined as follows: the action rules of the low and high type customers are given by

$$a_{H}(m) = \begin{cases} 
\text{join} & \text{if } m = m_{HL} \\
\text{balk} & \text{otherwise,}
\end{cases}$$

$$a_{L}(m) = \begin{cases} 
\text{join} & \text{if } m = m_{L} \text{ or } m = m_{HL} \\
\text{balk} & \text{otherwise.}
\end{cases}$$

In terms of the firm’s strategy, the firm provides three distinct messages $m_{\emptyset}$, $m_{L}$ and $m_{HL}$ which satisfy the condition given by (9). However, we cannot explicitly characterize the announcement policy. The optimal scheduling rule of the firm is given by

$$g(n_{H}, n_{L}, n_{HL}) = \begin{cases} 
m_{HL} & \text{if } n_{HL} > 0 \\
m_{L} & \text{if } n_{HL} = 0 \text{ and } n_{L} > 0 \\
m_{\emptyset} & \text{if } n_{HL} = n_{L} = 0
\end{cases}$$

with $n_{H} = 0$.

It is important to note that the equilibrium above requires the low type customers to be more patient than the high type customers, i.e., $\frac{R_{H}}{c_{H}} > \frac{R_{L}}{c_{L}}$. The question now is what if we have the low type customers to be more patient than the high type customers. Following a similar argument for the case with $h_{H} < h_{L}$ in Section 4.3, one can show that the firm achieves the best profit in a pooling equilibrium among all other equilibria, when we have $\frac{R_{H}}{c_{H}} > \frac{R_{L}}{c_{L}}$. The pooling equilibrium is identical to the one characterized in Proposition 2 but with the incentive compatibility condition given by $n + 2 > \frac{R_{HL}}{c_{HL}} > \frac{R_{L}}{c_{L}} \geq n + 1$ instead of (8).
Appendix B: Proofs

Proof of Theorem 1:

Given that there are two different actions, i.e., join and balk, for each customer type, there are four possible reactions from customers: all customers joining the system, the high type customers joining the system but not the low type customers, the low type customers joining the system but not the high type customers, and all customers balking. However, the second and the third reactions, i.e., the high type customers joining the system but not the low type customers, and the low type customers joining the system but not the high type customers, are mutually exclusive in equilibria. If there is an announcement $m$ which induces the outcome of the high type customers joining the system but not the low type customers, then we must have

$$R_H - c_H W_m > 0$$

and

$$R_L - c_L W_m < 0,$$

where $W_m$ is the expected waiting time of customers receiving the announcement $m$. Thus, we have $\frac{p_H}{c_H} > \frac{p_L}{c_L}$. However, if there is another announcement $m'$ which can induce the outcome of the low type customers joining the system but not the high type customers. Following similar arguments, we must have $\frac{p_L}{c_L} > \frac{p_H}{c_H}$, which leads to contradiction. Q.E.D.

Proof of Lemma 1:

We let $V(i,j)$ be the maximum expected total profit of the firm when there are $i$ high type and $j$ low type customers in the system. In order to characterize the first best solution of the firm, it is important to note that the optimality condition for the firm can be written as follows.

$$V(i,j) + \frac{\gamma t}{\lambda} = C(i,j) + \frac{\lambda_1}{\lambda} T_1 V(i,j) + \frac{\lambda_2}{\lambda} T_2 V(i,j) + \frac{\mu}{\lambda} T_3 V(i,j),$$

with

$$C(i,j) = -\frac{(h_L i + h_H j)}{\lambda},$$

$$T_1 V(i,j) = \max\{v_H + V(i+1,j), V(i,j)\},$$

$$T_2 V(i,j) = \max\{V(i,j+1) + v_L, V(i,j)\},$$

$$T_3 V(i,j) = \max\{V(i-1,j)I_{i>0} + V(i,j)I_{i=0}, V(i,j-1)I_{j>0} + V(i,j)I_{j=0}, V(i,j)\}.$$

$\lambda_1 = \beta_H \lambda$ and $\lambda_2 = \beta_L \lambda$. We next show that the optimal value function $V(i,j)$ is in $\bar{V}$, which is a set of functions defined as follows.
Definition 7. We define $\bar{V}$ as the set of functions such that if $V \in \bar{V}$, then $V$ satisfies the following conditions:

\begin{align*}
V(i, j) & \geq V(i + 1, j) \quad (14) \\
V(i, j) & \geq V(i, j + 1) \quad (15) \\
V(i, j + 1) + V(i + 1, j) & \geq V(i, j) + V(i + 1, j + 1) \quad (16) \\
V(i, j + 1) + V(i + 1, j + 1) & \geq V(i + 1, j) + V(i, j + 2) \quad (17) \\
V(i + 1, j) + V(i, j + 1) & \geq V(i, j + 1) + V(i + 2, j) \quad (18) \\
V(i, j + 1) & \geq V(i + 1, j) \text{ if } h_H > h_L; \\
V(i, j + 1) & \leq V(i + 1, j) \text{ if } h_H < h_L; \\
V(i, j + 1) & = V(i + 1, j) \text{ if } h_H = h_L.
\end{align*}

Before we show $V \in \bar{V}$, we first prove the following three lemmas, i.e., Lemma 3, 4 and 5. For exposition purposes, we present the proofs for Lemma 3, 4 and 5 at the end of the proof of this Proposition.

Lemma 3. If $V \in \bar{V}$, then $T_1 V \in \bar{V}$.

Lemma 4. If $V \in \bar{V}$, then $T_2 V \in \bar{V}$.

Lemma 5. If $V \in \bar{V}$, then $T_3 V \in \bar{V}$.

We are ready to show $V \in \bar{V}$. Consider a value iteration algorithm to solve for the optimal policy in which $V_0(i, j) = 0$ for all $i$ and $j$, and

$$V_{k+1}(i, j) = C(i, j) + \frac{\lambda_1}{\Lambda} T_1 V_k(i, j) + \frac{\lambda_2}{\Lambda} T_2 V_k(i, j) + \frac{\mu}{\Lambda} T_3 V_k(i, j)$$  \quad (20)

Based on Proposition 4.1.7 in Bertsekas et al. (2012), we have $\lim_{k \to \infty} V_k = V$. Thus, to show $V \in \bar{V}$, we only need to show $V_k \in \bar{V}$ for any $k \in \mathbb{Z}$. We do so by induction. Given that $V_0(i, j) = 0, \forall i, j \in \mathbb{Z}$, one should see $V_0 \in \bar{V}$. We now show if $V_k \in \bar{V}$, we have $V_{k+1} \in \bar{V}$. Based on Lemma 3, 4 and 5, if $V_k \in \bar{V}$, we have $T_1 V_k(i, j) \in \bar{V}$, $T_2 V_k(i, j) \in \bar{V}$ and $T_3 V_k(i, j) \in \bar{V}$. One should also see that $C(i, j) \in \bar{V}$. To this end, we have $V_{k+1} \in \bar{V}$ if $V_k \in \bar{V}$. Hence, by induction, we have $V_k \in \bar{V}$ for all $k \in \mathbb{Z}$. Given $\lim_{k \to \infty} V_k = V$, we have $V \in \bar{V}$.

Let us get back to the question of the firm’s optimal admission policy. We know that it is optimal for the firm to accept the high type customers when we have $V(i + 1, j) - V(i, j) > -v_H$. Due to $V \in \bar{V}$, one should see that $V(i + 1, j) - V(i, j)$ is a non-increasing function in $j$ based on property (16). Moreover, based on (16)+(18), one can see that $V(i + 1, j) - V(i, j)$ is a non-increasing function in $i$. To this end, one can show that the firm’s optimal admission policy to the high type customers can be characterized by a finite switching curve $S_H(j)$ defined as follows

$$S_H(j) = \max\{i : V(i + 1, j) - V(i, j) > -v_H \mid i, j \in \mathbb{Z}\},$$
where \( i \) is the number of high type customers in the system and \( j \) is the number of low type customers. Similarly, one can show that the firm’s optimal admission policy to the low type customers can be characterized by a finite switching curve \( S_L(i) \) defined as follows
\[
S_L(i) = \max \{ j : V(i, j + 1) - V(i, j) > -v_L, i \in \mathbb{Z} \}.
\]

As for the firm’s optimal scheduling policy, based on (14), (15) and (19), one should see that, when we have \( h_H \neq h_L \), it is optimal for the firm to give preemptive resume priority to customers of type \( k \) in the system, where \( k \) is given by
\[
k = \arg \max_{i \in \{H, L\}} \{ h_i \}.
\]
When we have \( h_H = h_L \), the order of service does not impact the profit of the firm. (Please see the proofs for Lemma 3, 4 and 5 as follows.) Q.E.D.

**Proof of Lemma 3:**

To show \( T_1 V_k(i, j) \in V \) if \( V_k(i, j) \in V \), we show the following:
- We next show \( T_1 \) preserves the properties given by (14). We let \( y \) denote the optimal action for the firm in the state \((i + 1, j)\). In particular, \( y = 0 \) means that it is optimal for the firm to reject the high type customer when the system state is \((i + 1, j)\), while \( y = 1 \) means that it is optimal for the firm to accept the high type customer:
  - when \( y = 0 \), we have
    \[
    T_1 V_k(i, j) = \max \{ v_H + V_k(i + 1, j), V_k(i, j) \} \\
    \geq V_k(i, j) \\
    \geq V_k(i + 1, j) = T_1 V_k(i + 1, j),
    \]
    where the second inequality is based on the condition given by (14).
  - Similar, when \( y = 1 \), we have
    \[
    T_1 V_k(i, j) = \max \{ v_H + V_k(i + 1, j), V_k(i, j) \} \\
    \geq v_H + V_k(i + 1, j) \\
    \geq v_H + V_k(i + 2, j) = T_1 V_k(i + 1, j)
    \]
    Thus, we have shown that the operator \( T_1 \) preserves the property given by (14).
- We next show that \( T_1 \) preserves the property given by (15). Similarly, we let \( y \) denote the optimal action for the firm in the state \((i, j + 1)\). In particular, \( y = 0 \) means that it is optimal for the firm to reject the high type customer when the system state is \((i, j + 1)\), while \( y = 1 \) means that it is optimal for the firm to accept the high type customer:
  - when \( y = 0 \), we have
    \[
    T_1 V_k(i, j) = \max \{ v_H + V_k(i + 1, j), V_k(i, j) \} \\
    \geq V_k(i, j) \\
    \geq V_k(i, j + 1) = T_1 V_k(i, j + 1),
    \]
    where the second inequality is based on the condition given by (15).
— Similar, when \( y = 1 \), we have

\[
T_1 V_k(i, j) = \max\{v_H + V_k(i + 1, j), V_k(i, j)\}
\]

\[
\geq v_H + V_k(i + 1, j)
\]

\[
\geq v_H + V_k(i + 1, j + 1)
\]

\[
= T_1 V_k(i, j + 1)
\]

— We now show that \( T_1 \) preserves the property given by (16). Similarly, we let \( y_1 \) and \( y_2 \) denote the optimal action for the firm in the state \((i, j)\) and \((i+1, j+1)\). In particular, \( y_1 = 0 \) means that it is optimal for the firm to reject the high type customer when the system state is \((i, j)\), accept otherwise. Moreover, \( y_2 = 0 \) means that it is optimal for the firm to reject the high type customer when the system state is \((i+1, j+1)\), accept otherwise:

— When we have \( y_1 = y_2 = 0 \),

\[
T_1 V_k(i, j + 1) + T_1 V_k(i + 1, j) \geq V_k(i, j + 1) + V_k(i + 1, j)
\]

\[
\geq V_k(i, j) + V_k(i + 1, j + 1)
\]

\[
= T_1 V_k(i, j) + T_1 V_k(i + 1, j + 1),
\]

where the second inequality is based on the condition given by (16).

— When we have \( y_1 = 1 \) and \( y_2 = 0 \),

\[
T_1 V_k(i, j + 1) + T_1 V_k(i + 1, j) \geq v_H + V_k(i + 1, j + 1) + V_k(i + 1, j)
\]

\[
= T_1 V_k(i, j) + T_1 V_k(i + 1, j + 1),
\]

— When we have \( y_1 = 0 \) and \( y_2 = 1 \), we show below it leads to contradiction. Given that \( y_1 = 0 \), we have

\[
V_k(i, j) - V_k(i + 1, j) \geq v_H;
\]

Similarly, given that we have \( y_2 = 1 \), hence, \( V_k(i + 1, j + 1) - V(i + 2, j + 1) \leq v_H \). Therefore, we have

\[
V_k(i, j) + V_k(i + 2, j + 1) \geq V_k(i + 1, j + 1) + V_k(i + 1, j)
\]

(21)

However, it is important to note that we have (16) with \( i \) replaced by \( i + 1 \), (16) and (18), hence, we have

\[
V_k(i, j) + V_k(i + 2, j + 1) \leq V_k(i + 1, j + 1) + V_k(i + 1, j).
\]

This contradict to (21) above.

— When we have \( y_1 = y_2 = 1 \),

\[
T_1 V_k(i, j + 1) + T_1 V_k(i + 1, j) \geq v_H + V_k(i + 1, j + 1) + v_H + V_k(i + 2, j)
\]

\[
\geq v_H + V_k(i + 1, j) + v_H + V_k(i + 2, j + 1)
\]

\[
= T_1 V_k(i, j) + T_1 V_k(i + 1, j + 1),
\]

where the second inequality is based on the condition given by (16) with \( i \) replaced by \( i + 1 \).

— We now show that \( T_1 \) preserves the property given by (17). Similarly, we let \( y_1 \) and \( y_2 \) denote the optimal action for the firm in the state \((i + 1, j)\) and \((i, j + 2)\). In particular, \( y_1 = 0 \) means that it is optimal for the firm to reject the high type customer when the system state is \((i + 1, j)\), accept otherwise. Moreover, \( y_2 = 0 \) means that it is optimal for the firm to reject the high type customer when the system state is \((i, j + 2)\), accept otherwise:
— When we have \( y_1 = y_2 = 0 \),
\[
T_1 V_k(i, j + 1) + T_1 V_k(i + 1, j + 1) \geq V_k(i, j + 1) + V_k(i + 1, j + 1)
\]
\[
\geq V_k(i + 1, j) + V_k(i, j + 2)
\]
\[
= T_1 V_k(i + 1, j) + T_1 V_k(i, j + 2),
\]
where the second inequality is based on (17).

— When we have \( y_1 = y_2 = 1 \),
\[
T_1 V_k(i, j + 1) + T_1 V_k(i + 1, j + 1) \geq v_H + V_k(i + 1, j + 1) + v_H + V_k(i + 2, j + 1)
\]
\[
\geq v_H + V_k(i + 2, j) + v_H + V_k(i + 1, j + 2)
\]
\[
= T_1 V_k(i + 1, j) + T_1 V_k(i, j + 2),
\]
where the second inequality is based on (17).

— When we have \( y_1 = 1 \) and \( y_2 = 0 \),
\[
T_1 V_k(i, j + 1) + T_1 V_k(i + 1, j + 1) \geq V_k(i + 1, j + 1) + v_H + V_k(i + 1, j + 1)
\]
\[
\geq V_k(i + 1, j) + v_H + V_k(i, j + 2)
\]
\[
= T_1 V_k(i + 1, j) + T_1 V_k(i, j + 2),
\]
where the second inequality is based on the summation of (17) and (18).

— When we have \( y_1 = 0 \) and \( y_2 = 1 \),
\[
T_1 V_k(i, j + 1) + T_1 V_k(i + 1, j + 1) \geq V_k(i + 1, j + 1) + v_H + V_k(i + 1, j + 1)
\]
\[
\geq V_k(i + 1, j) + v_H + V_k(i, j + 2)
\]
\[
= T_1 V_k(i + 1, j) + T_1 V_k(i, j + 2),
\]
where the second inequality is based on the summation of (16) and (17).

• We now show that \( T_1 \) preserves the property given by (18). Similarly, we let \( y_1 \) and \( y_2 \) denote the optimal action for the firm in the state \((i, j + 1)\) and \((i + 2, j)\). In particular, \( y_1 = 0 \) means that it is optimal for the firm to reject the high type customer when the system state is \((i, j + 1)\), accept otherwise. Moreover, \( y_2 = 0 \) means that it is optimal for the firm to reject the high type customer when the system state is \((i + 2, j)\), accept otherwise:

— When we have \( y_1 = y_2 = 0 \),
\[
T_1 V_k(i + 1, j) + T_1 V_k(i + 1, j + 1) \geq V_k(i + 1, j) + V_k(i + 1, j + 1)
\]
\[
\geq V_k(i, j + 1) + V_k(i + 2, j)
\]
\[
= T_1 V_k(i, j + 1) + T_1 V_k(i + 2, j),
\]
where the second inequality is due to (18).
— When we have \( y_1 = 1 \) and \( y_2 = 0 \),

\[
T_1 V_k(i + 1, j) + T_1 V_k(i + 1, j + 1) \geq V_k(i + 2, j) + v_H + V_k(i + 1, j + 1) = T_1 V_k(i + 2, j) + T_1 V_k(i, j + 1)
\]

— When we have \( y_1 = 0 \) and \( y_2 = 1 \), we show that it is not feasible. Given that we have \( y_1 = 0 \) and \( y_2 = 1 \), we get

\[
V_k(i, j + 1) + V_k(i + 3, j) \geq V_k(i + 1, j + 1) + V_k(i + 2, j) \tag{22}
\]

To this end, it is important to note that by replacing \( i \) with \( i + 1 \) in (18), we get

\[
V_k(i + 2, j) + V_k(i + 2, j + 1) \geq V_k(i + 1, j + 1) + V_k(i + 3, j).
\]

Similarly, by replacing \( i \) with \( i + 1 \) in (16), we get

\[
V_k(i + 1, j + 1) + V_k(i + 2, j) \geq V_k(i + 1, j) + V_k(i + 2, j + 1).
\]

Summing up the above two inequalities together with (18), we get

\[
V_k(i + 2, j) + V_k(i + 1, j + 1) \geq V_k(i + 3, j) + V_k(i, j + 1),
\]

which contradicts to (22).

— When we have \( y_1 = y_2 = 1 \), the proof is similar to the case when we have \( y_1 = y_2 = 0 \).

- We now show that \( T_1 \) preserves the property given by (19). Similarly, we let \( y_1 \) denote the optimal action for the firm in the state \((i + 1, j)\). In particular, \( y_1 = 0 \) means that it is optimal for the firm to reject the customer when the system state is \((i + 1, j)\), accept otherwise. Below, we start with the case \( h_H > h_L \), while the cases when \( h_H \leq h_L \) can be shown in a similar manner.

— When we have \( y_1 = 0 \),

\[
T_1 V_k(i, j + 1) \geq V_k(i, j + 1) \geq V_k(i + 1, j) = T_1 V_k(i + 1, j)
\]

— When we have \( y_1 = 1 \),

\[
T_1 V_k(i, j + 1) \geq v_H + V_k(i + 1, j + 1) \geq V_k(i + 2, j) + v_H = T_1 V_k(i + 1, j)
\]

It is important to note that we have only used property given in (19) to show that \( T_1 \) preserves the property given by (19). This implies that the optimal priority policy solely depends on the per unit holding cost of each of the customer type regardless of the announcement policy.

Thus, we have proved Lemma 3. Q.E.D.

**Proof of Lemma 4:**

The proof is similar to the proof of Lemma 3 above. Q.E.D.

**Proof of Lemma 5:**

We start with the proof for the case when we have \( h_H > h_L \). Note that since \( V_k \in \bar{V} \), so when \( h_H > h_L \), \( T_3 V_k(i, j) \) is equivalent to

\[
T_3 V_k(i, j) = V_k(i - 1, j)I_{i \geq 1} + V_k(0, j - 1)I_{i = 0, j \geq 1} + V_k(0, 0)I_{i = j = 0}
\]

- We now show that \( T_3 \) preserves the property given by (14). If \( i \geq 1 \) and \( j \geq 0 \), we have \( T_3 V_k(i, j) = V_k(i - 1, j) \geq V_k(i, j) = T_3(i + 1, j) \); If \( i = 0 \) and \( j \geq 1 \), \( T_3 V_k(i, j) = V_k(i, j - 1) \geq V_k(i, j) = T_3 V_k(i + 1, j) \); And if \( i = j = 0 \), \( T_3 V_k(0, 0) = V_k(0, 0) = T_3 V_k(1, 0) \).
• We now show that \( T_3 \) preserves the property given by (15). It is similar to the proof above.

• We now show that \( T_3 \) preserves the property given by (16), i.e., \( T_3 V_k(i, j + 1) + T_3 V_k(i + 1, j) \geq T_3 V_k(i, j) + T_3 V_k(i + 1, j + 1) \).
  
  — if \( i \geq 1 \) and \( j \geq 0 \),
  
  \[
  T_3 V_k(i, j + 1) + T_3 V_k(i + 1, j) = V_k(i - 1, j + 1) + V_k(i, j) \\
  \geq V_k(i - 1, j) + V_k(i, j + 1) = T_3 V_k(i, j) + T_3 V_k(i + 1, j + 1);
  \]

  — if \( i = 0 \) and \( j \geq 0 \),
  
  \[
  T_3 V_k(i, j + 1) + T_3 V_k(i + 1, j) = V_k(0, j) + V_k(0, j) \\
  \geq V_k(0, j - 1) + V_k(0, j + 1) = T_3 V_k(i, j) + T_3 V_k(i + 1, j + 1);
  \]

where the inequality is based on condition given by summation of (16) and (17).

• We now show that \( T_3 \) preserves the property given by (17), i.e., \( T_3 V_k(i, j + 1) + T_3 V_k(i + 1, j + 1) \geq T_3 V_k(i + 1, j) + T_3 V_k(i, j) \).

  — if \( i \geq 1 \) and \( j \geq 0 \),
  
  \[
  T_3 V_k(i, j + 1) + T_3 V_k(i + 1, j + 1) = V_k(i - 1, j + 1) + V_k(i, j + 1) \\
  \geq V_k(i, j) + V_k(i - 1, j + 2) = T_3 V_k(i + 1, j) + T_3 V_k(i, j + 2),
  \]

where the inequality is due to (17).

  — if \( i = 0 \) and \( j \geq 0 \),
  
  \[
  T_3 V_k(0, j + 1) + T_3 V_k(1, j + 1) = V_k(0, j) + V_k(0, j + 1) \\
  = T_3 V_k(1, j) + T_3 V_k(0, j + 2),
  \]

• We now show that \( T_3 \) preserves the property given by (18), i.e., \( T_3 V_k(i + 1, j) + T_3 V_k(i + 1, j + 1) \geq T_3 V_k(i, j + 1) + T_3 V_k(i + 2, j) \).

  — if \( i \geq 1 \) and \( j \geq 0 \),
  
  \[
  T_3 V_k(i + 1, j) + T_3 V_k(i + 1, j + 1) = V_k(i, j) + V_k(i, j + 1) \\
  \geq V_k(i - 1, j + 1) + V_k(i + 1, j) = T_3 V_k(i, j + 1) + T_3 V_k(i + 2, j),
  \]

where the inequality is due to (18).

  — if \( i = 0 \) and \( j \geq 0 \),
  
  \[
  T_3 V_k(1, j) + T_3 V_k(1, j + 1) = V_k(0, j) + V_k(0, j + 1) \\
  \geq V_k(0, j) + V_k(1, j) = T_3 V_k(0, j + 1) + T_3 V_k(2, j),
  \]

where the inequality is due to (19).

• We now show that \( T_3 \) preserves the property given by (19), i.e., \( T_3 V_k(i, j + 1) \geq T_3 V_k(i + 1, j) \), assuming \( h_H > h_L \). If \( i \geq 1 \), we have \( T_3 V_k(i, j + 1) = V_k(i - 1, j + 1) \geq V(i, j) = T_3 V_k(i + 1, j) \), where the second equality is due to (19); If \( i = 0 \), we have \( T_3 V_k(0, j + 1) = V_k(0, j) = T_3 V_k(1, j) \).

We have shown the case when \( h_H > h_L \). The cases when \( h_H < h_L \) and \( h_H = h_L \) can be shown in a similar manner, Q.E.D.
Proof of Lemma 2:

We know that the switching curves $S_H(.)$ and $S_L(.)$ given in Proposition 1 are defined as follows:

$$S_H(j) = \max\{i : V(i + 1, j) - V(i, j) > -v_H| i, j \in \mathbb{Z}\}$$

$$S_L(i) = \max\{j : V(i, j + 1) - V(i, j) > -v_L| i, j \in \mathbb{Z}\}$$

We let $S_H(0) = \hat{n}_H^L$, to show $S_H(j) = \hat{n}_H^L - j$, we only need to show $S_H(j + 1) = S_H(j) - 1$. We know

$$S_H(j + 1) = \max\{i : V(i + 1, j + 1) - V(i, j + 1) > -v_H| i, j \in \mathbb{Z}\}$$

$$= \max\{i : V(i + 2, j) - V(i + 1, j) > -v_H| i, j \in \mathbb{Z}\}$$

$$= \max\{i' - 1 : V(i' + 1, j) - V(i', j) > -v_H| i', j \in \mathbb{Z}\}$$

$$= \max\{i : V(i + 1, j) - V(i, j) > -v_H| i, j \in \mathbb{Z}\} - 1$$

$$= S_H(j) - 1$$

The second equality is due to the property $V(i + 1, j) = V(i, j + 1)$, see (19) in the proof of Proposition 1. Thus, we have shown $S_H(n_H^L) = \hat{n}_H^L - n_H^L$. Similarly, we let $S_L(0) = \hat{n}_L^H$, we then can show $S_L(n_L^H) = \hat{n}_L^H - n_L^H$.

Meanwhile, we have

$$S_H(0) = \max\{i| V(i + 1, 0) - V(i, 0) > -v_H| i, j \in \mathbb{Z}\}$$

$$S_L(0) = \max\{j| V(0, j + 1) - V(0, j) > -v_L| i, j \in \mathbb{Z}\} = \max\{j| V(j + 1, 0) - V(j, 0) > -v_L| i, j \in \mathbb{Z}\}$$

As it is shown in the proof of proposition 1, $V(j + 1, 0) - V(j, 0)$ is decreasing in $j$. To this end, we have $\hat{n}_H^L \geq \hat{n}_L^H$ if $v_H > v_L$. Q.E.D.

Proof of Proposition 1:

It is clear that the proposed equilibrium achieves the first best for the firm and hence the firm does not have any profitable deviation. For the customer, one can see that if the message provided is $m_1$, the number of customers in the system is $\bar{n}_L$. Hence, the average waiting time experienced by the customers who join the system when the firm announce $m_1$ is $\frac{\bar{n}_L + 1}{\mu}$. Based on the (5) and (6) given in the proposition, customers of both types are better off by joining the system when the announcement received is $m_1$. With similar arguments, one can show that only high type customers are better off by joining the system when the announcement received is $m_2$, while both high type and low type customers are better off to balk when the announcement received is $m_3$. Q.E.D.

Proof of Proposition 2:

The proof of $(a_L, a_H, A, g)$ is an equilibrium is similar to the proof of Proposition 1 above. We next show that there does not exist any equilibrium which obtains a higher profit than $(a_L, a_H, A, g)$ characterized in the proposition. Note that under any equilibrium $(a_L', a_H', A', g')$, given $\frac{n_H}{c_H} < \frac{n_L}{c_L}$, we have $a_L'(m) \geq a_H'(m)$ for all $m$ that are used with positive probability in the equilibrium. To this end, let $\pi$ denote the profit of
the firm when it cannot observe customer type and take the following actions: (1) allow both customers to join the system; (2) allow only low type customers to join; and (3) allow neither type of customers to join. It’s worth mentioning that allowing only high type customers to join can not be sustained in any equilibria. Moreover, given \( v_H > v_L \) and \( h_H = h_L \), we obtain that it is never optimal for the firm to allow only the low type customers to join. Thus, \( \pi \) is the same as the profit of the firm when it treats customers of both types identically. Hence, the firm’s profit is bounded by \( \pi \) when it does not observe customer types. Q.E.D.

Proof of Theorem 2:

Based on Proposition 1, we show that, when the per unit holding cost is different for customers of different types, to achieve the first, the firm should give absolute priority to the type of customers with a relatively higher per unit holding cost between the two types of customers. However, the firm cannot directly observe the type of customers. As a result, it can only prioritize the customers whose types it elicits based on their responses towards the announcements. Based on Proposition 1, one can see that, to achieve the first best, the firm would like to admit both customer types when there are no customers in the system, for any non-degenerate case with \( S_i(0) \geq 0, \forall i \in \{H,L\} \). As a result, to achieve the first best, the firm must provide at least one message which induces both customer types to join the system. The firm cannot differentiate the customers who receive such an announcement in the system. Hence, the firm cannot prioritize these customers appropriately which prevents the firm from achieving the first best.

Proof of Proposition 3:

We start with the firm’s optimal strategy, which is comprised of the announcement policy and the priority policy. Note that the firm’s optimal policy can be characterized by the following optimality equation.

\[
V(i, j, k) + \gamma \Lambda = \bar{C}(i, j, k) + \frac{\lambda}{\Lambda} T_4 V(i, j, k) + \frac{\mu}{\Lambda} T_5 V(i, j, k),
\]

with

\[
\bar{C}(i, j, k) = \frac{- (h_H \beta_H + h_L \beta_L) k - h_H i - h_L j}{\Lambda}
\]

\[
T_4 V(i, j) = \max_{m \in M} \left\{ (V(i, j, k + 1) + \beta_H v_H + \beta_L v_L) I_{(m \in M(H,L))},
(\beta_H V(i + 1, j, k) + \beta_L V(i, j, k) + \beta_H v_H) I_{(m \in M(H))},
(\beta_L V(i, j + 1, k) + \beta_H V(i, j, k) + \beta_L v_L) I_{(m \in M(L))},
V(i, j, k) I_{(m \in S_k)} \right\}.
\]

and

\[
T_5 V(i, j) = \max \left\{ V(i - 1, j, k) I_{(i > 0)} + V(i, j, k) I_{(i = 0)},
V(i, j - 1, k) I_{(j > 0)} + V(i, j, k) I_{(j = 0)},
V(i, j, k - 1) I_{(k > 0)} + V(i, j, k) I_{(k = 0)},
V(i, j, k) \right\},
\]

where \( i, j, k \) are the numbers of customers receiving message \( m_1 \in M(H), m_3 \in M(L) \) and \( m_2 \in M(H,L) \), respectively.
We next show that the optimal priority policy of the firm is given by

\[
g(i, j, k) = \begin{cases} 
m_3 & \text{if } j > 0 \\
m_2 & \text{if } j = 0 \text{ and } k > 0 \\
m_1 & \text{if } j = k = 0 \text{ and } i > 0 \\
m_0 & \text{if } j = k = i = 0,
\end{cases}
\]  

(24)

with \(m_0 \in M_0\). Before, we start the proof, we define the set of function \(G\) as follows.

**Definition 8.** If a function \(V \in G\), then the function \(V\) satisfies the following properties:

\[
V(i, j, k) \geq V(i + 1, j, k) \\
V(i + 1, j, k) \geq V(i, j + 1, k).
\]  

(25)

(26)

Note that with probability \(\beta_i\) with \(i \in \{H, L\}\), a customer receiving message \(m_2\) is a type \(i\) customer. Thus, we have \(V(i, j, k + 1) = \beta_H V(i + 1, j, k) + \beta_L V(i, j + 1, k)\). As a result, the condition \(V(i + 1, j, k) \geq V(i, j + 1, k)\) is equivalent to \(V(i + 1, j, k) \geq V(i + 1, j, k + 1) \geq V(i, j + 1, k)\). To this end, to show the optimal priority policy is given by (24), it is equivalent to show that the value function of the firm \(V \in G\). In order to show that \(V \in G\), following a similar logic to the one used in the proof for Proposition 1, it is sufficient to show the following two lemmas.

**Lemma 6.** if \(V \in G\), then \(T_4V \in G\).

**Lemma 7.** if \(V \in G\), then \(T_5V \in G\).

- We now start proving Lemma 6:

  We next show that \(T_4\) preserves the property characterized by (26), which is equivalent to show that if \(V \in G\), then \(T_4V(i + 1, j, k) \geq T_4V(i, j + 1, k)\). In order to do so, we let \(m\) represent the optimal action of the firm when the system state is \((i, j + 1, k)\). If \(m \in M_{(H,L)}\), we have \(T_4V(i + 1, j, k) \geq V(i + 1, j, k + 1) + \beta_H v_H + (1 - \beta_H) v_L \geq V(i, j + 1, k + 1) + \beta_H v_H + (1 - \beta_H) v_L = T_4V(i, j + 1, k)\); when \(m \in M_{(H)}\), we have \(T_4V(i + 1, j, k) \geq \beta_H V(i + 2, j, k) + \beta_H v_H + (1 - \beta_H) V(i + 1, j, k) \geq \beta_H V(i + 1, j, k + 1) + \beta_H v_H + (1 - \beta_H) V(i, j + 1, k) = T_4V(i, j + 1, k)\); When \(m \in M_{(L)}\), we have

  \[
  T_4V(i + 1, j, k) \geq \beta_H V(i + 1, j, k) + \beta_L V(i + 1, j + 1, k) + \beta_L V_L \\
  \geq \beta_L V(i, j + 2, k) + \beta_L V_L + \beta_H V_L(i, j + 1, k) \\
  = T_4V(i, j + 1, k);
  \]

  When \(m \in M_0\), we have \(T_4V(i + 1, j, k) \geq V(i + 1, j, k) \geq V(i, j + 1, k) = T_4V(i, j + 1, k)\). To this end, we have shown that if \(V \in G\), then \(T_4V\) satisfies condition (26).

  The proof for that \(T_4\) preserves the property given in (25) is similar to the one above.

- We next prove Lemma 7. We start by showing that \(T_5\) preserves the property characterized by (26), which is equivalent to show that if \(V \in G\), then \(T_5V(i + 1, j, k) \geq T_5V(i, j + 1, k)\). In order to do so, we let \(m\) represent the optimal announcement to provide for the firm when the system state is \((i, j + 1, k)\). If \(j > 0\), we have \(T_5V(i + 1, j, k) = V(i + 1, j - 1, k) \geq V(i, j, k) = T_5V(i, j, k)\); If \(j = 0\) and \(k > 0\), we have \(T_3V(i + 1, j, k) = V(i + 1, j, k - 1) \geq V(i, j, k) = T_3V(i, j, k)\); If \(j = k = 0\), \(T_5V(i + 1, j, k) = V(i, j, k) \geq T_5V(i, j + 1, k)\). To this end, we have shown \(T_5\) preserves the property (26). The proof for that \(T_5\) preserves property (25) is similar.
Based on the proof above, we have shown that the optimal priority policy of the firm is given by (24). To this end, \( T_3 V(i,j,k) \) defined in the optimality condition (23) can be simplified to be

\[
T_3 V(i,j,k) = V(i,j-1,k)I_{j \geq 1} + V(i,j,k-1)I_{j=0,k \geq 1} + V(i-1,j,k)I_{j=k=0, i > 0} + V(0,0,0)I_{i=j=k=0}.
\]

We next show that, when we have \( h_H < h_L \), it is never optimal for the firm to provide message \( m \in M_{(L)} \).

Let us first assume that there exists \((i,j,k)\) such that an announcement \( m \in M_{(L)} \) is the optimal one to provide. To this end, we have

\[
\beta_L V(i,j+1,k) + \beta_H V(i,j,k) + \beta_L v_L > V(i,j,k+1) + \beta_H v_H + \beta_L v_L
\]

and

\[
\beta_L V(i,j+1,k) + \beta_H V(i,j,k) + \beta_L v_L > V(i,j,k).
\]

Given that we have \( V(i,j,k+1) = \beta_H V(i+1,j,k) + \beta_L V(i,j+1,k) \) together with the above two inequalities, we have

\[
V(i,j,k) - V(i+1,j,k) > v_H
\]

and

\[
v_L > V(i,j,k) - V(i,j,k+1)
\]

Given that we have \( V(i,j+1,k) \leq V(i+1,j,k) \) and \( v_H > v_L \), (27) and (28) contradict to each other. Thus, there exist no states such that a message \( m \in M_{(L)} \) is an optimal announcement to provide.

Recall that we focus on the non-degenerate cases where it is optimal for the firm to admit customers of both types when there are no customers in the system. Thus, in equilibria, there must exist a message \( m_{2} \in M_{(HL)} \) which induces both customer types to join the system. Given that we have \( v_H > v_L \) and \( h_H < h_L \), the firm may like to provide a message with \( m_1 \in M_{(H)} \) to induce high type customers to join but low type customers to balk. Meanwhile, given \( h_i > 0 \) for \( i \in \{H,L\} \), in any equilibria, the firm would like to provide a message with \( m_2 \in M_{H} \) to induce both customer types to balk when the system is really congested. Above, we have shown that there is no incentive for the firm to deviate. As for the customers, given incentive compatibility conditions given in (9), it is better off for both customer types to join when they receive message \( m_1 \), while it is better off for high type customers but not low type customers to join when they receive message \( m_2 \). It is better off for both customer types to balk when they receive message \( m_0 \). Q.E.D.

**Proof of Proposition 4:**

We start by exploring the conditions when the babbling equilibrium where both types of the customers join the system regardless of the announcements may exist. If customers of both types indeed join the queue disregard of the announcements received, the system becomes an \( M/M/1 \) system with the arrival rate and the service rate being \( \lambda \) and \( \mu \), respectively. Given that the firm cannot differentiate customer types in any way through a babbling equilibrium, we focus on the case when the firm serves the customers in a first-come, first-served manner. Thus, one can show that the average waiting time in the system is given by \( \frac{1}{\mu - \lambda} \). Since customers would join the system if and only if their expected utility is positive in equilibrium, we have \( R_i - \frac{1}{\mu - \lambda} \geq 0, \forall i \in \{H,L\} \). Following a similar logic, we can characterize the other two types of babbling equilibria as described in Proposition 4. Q.E.D.
Proof of Proposition 5:

**Firm’s Profit:** We let \( \Pi_{IP} \) be the profit of the firm per unit time under the influential pooling equilibrium, while let \( U_{IP} \) be the utility of the customers per unit of time. We let the firm’s profit per unit time under the system \( M/M/1/k \) be \( \Omega(k) \). Based on Theory 1 in Knudsen (1972), \( \Omega(k) \) is a unimodal in \( k \). In particular, there exist a finite \( k^* \in \mathbb{Z}_+ \) such that the function \( \Omega(k) \) is strictly increasing for \( k < k^* \) and is strictly decreasing for \( k \geq k^* \). To have the pooling equilibrium hold, we have \( k^* = \hat{n}^f + 1 \). Meanwhile, the system under the babbling equilibrium where both customer types join is equivalent to \( M/M/1/\infty \). Thus, the firm’s profit under the pooling equilibrium is larger than the firm’s profit under the babbling equilibrium, i.e., \( \Pi_{IP} > \Pi_{NI} \).

**Customer Utility:** Recall that the system dynamic under a pooling equilibrium is the same as \( M/M/1/\hat{n}^f + 1 \), where there is only one customer type. In particular, for these customers, the value obtained by the firm through serving each customer, the per unit holding cost, the reward of service for the customers and the per unit waiting cost of the customers are given by \( \beta_h v_H + \beta_L v_L, \beta_H h_H + \beta_L h_L, \beta_H R_H + \beta_L R_L, \) and \( \beta_H c_H + \beta_L c_L \), respectively. To this end, under the full information case where customers can observe the number of customers in the system, the threshold for these customers are given by \( n^*_{HL} = \left\lfloor \frac{\beta_H R_H + \beta_L R_L}{\beta_H v_H + \beta_L v_L} \right\rfloor \).

We now let customers overall utility per unit time under the system \( M/M/1/k \) be \( \Omega^*(k) \). Based on the results in Section 4 of Naor (1969), there exists \( k^* \in \mathbb{Z}_+ \) such that the function \( \Omega^*(k) \) is strictly increasing for \( k < k^* \) and is strictly decreasing for \( k \geq k^* \). Naor (1969) also shows \( k^* < n^*_{HL} \). Meanwhile, one can show that \( n^*_{HL} \leq \max\{n^*_H, n^*_L\} \) with \( n^*_H = \left\lfloor \frac{\beta_H h_H}{\beta_H v_H} \right\rfloor \) and \( n^*_L = \left\lfloor \frac{\beta_L h_L}{\beta_L v_L} \right\rfloor \). Moreover, we have \( \hat{n}^f > \max\{n^*_H, n^*_L\} \) in order to have the pooling equilibrium to hold. Thus, we have \( k^* < \hat{n}^f \). As a result, the overall customers utility under the pooling equilibrium is larger than the one in the system under the babbling equilibrium where the threshold is equivalent to be \( \infty \), i.e., \( U^*_{IP} > U^*_{NI} \). Q.E.D.

Proof of Theorem 3:

We start from the case when the holding cost is the same for both customers types, i.e., \( h_H = h_L \). Recall that Proposition 2 shows that, to achieve the first best, the firm would like both types of customers to join the system when the number of customers in the system is smaller than \( \hat{n}^f_L \), would like high type customers to join but not the low type when the number of customers is between \( \hat{n}^f_L \) and \( \hat{n}^f_H \), and would like both customer types to balk otherwise. To this end, when the firm observes the type of the customers, for any influential equilibrium to exist, the only threshold for the low type customers which immunes from profitable deviations by the firm is \( \hat{n}^f_L \). Similarly, one can show that \( \hat{n}^f_H \) is the only threshold for the high type customers which prevents the firm from profitable deviations. To this end, we have shown that, assuming \( h_H = h_L \), under any MPBNE with influential cheap talk, the firm achieves the first best. Similar arguments apply for the case when we have \( h_H \neq h_L \). Q.E.D.

Proof of Proposition 6:

It is clear that the proposed equilibrium achieves the first best for the firm and hence the firm does not have any profitable deviations. For the high type customers, one can see that if the message provided is \( m^*_H \), the number of high type customers in the system denoted by \( n^0_H \) is larger than a threshold given by
$S_H(n^0_L)$. Hence, the expected waiting time of the arriving high type customer who receives announcement $m^H_2$ is given by $\bar{w}_H$, which is equivalent to

$$\sum_{n^0_L=0}^{\infty} \sum_{n^0_H=0}^{\infty} (R_H - c_H w_H^m(n^0_H, n^0_L))p(n^0_H, n^0_L | a_H, a_L, g_I, A^H, A^L) < 0$$

The equation above implies that the high type customers would obtain negative utility by joining the system when they receive the message $m^H_2$. Hence, it is better off for the high type customers to balk the system when they receive message $m^H_2$. Similarly, we can show that it is better off for the high type customers to join the system when they receive message $m^H_1$. Thus, high type customers would have no incentive to deviate from the equilibrium. Following a similar argument, we can show that the low type customers do not have incentive to deviate either. Q.E.D.

**Proof of Proposition 7:**

Note we have $\bar{w}_H = \frac{\nu_H}{\mu} + 2$ for the case with $h_H = h_L$. Thus, to show $D^I_N \cap D_I \neq \emptyset$, it is sufficient to show that $w_H < \frac{\bar{w}_H}{\mu} + 1$. We know $w_H = \frac{\mathbb{E}_{FB}[n|0 \leq n \leq \hat{n}_L]}{\mu} + 1$, while we have $\bar{n}_H = \mathbb{E}_{FB}[\hat{n}_L < n \leq \hat{n}_H]$. To this end, one can see $\bar{w}_H < \frac{\bar{w}_H}{\mu} + 1$.

When we have $h_H = h_L$, we also have $\bar{w}_L = \frac{\mathbb{E}_{FB}[n|\hat{n}_L < n \leq \hat{n}_H]}{\mu}$. Thus, by definition, we have $\bar{w}_L > \frac{\bar{w}_H}{\mu} + 1$.

Together with the result $w_H < \frac{\bar{w}_H}{\mu} + 1$, we have $D^{SS}_N \subset D_I$. Thus, to show $D^I_N \cap D_I = \emptyset$, it is equivalent to show $D^P_N \subset D_I$. It is trivial to see that $D^P_N \subset D_I$ is equivalent to $\mu w_H \leq \hat{n} + 1 \leq \hat{n} + 2 \leq \mu \bar{w}_L$. Q.E.D.

**Proof of Proposition 8:**

When we have $h_H < h_L$, the low type customers have the absolute priority over the high type customers. Thus, we have $\bar{w}_L < w_H$. To this end, $D^I_N \subset (D^I_I \cap D^I_N)$. We know $D^P_N \neq \emptyset$. Thus, we have $D^I_I \cap D^I_N \neq \emptyset$ when we have $h_H < h_L$.

$D^I_I \cap D_I = \emptyset$ is equivalent to $D_I \subset D^{SS}_N$. One can also see that $D_I \subset D^{SS}_N$ is equivalent to $\bar{w}_m \leq w_L \leq \bar{w}_L \leq \bar{w}_m \leq w_H \leq \bar{w}_H \leq \bar{w}_m$. Q.E.D.