

The impact of delaying the delay announcements

Gad Allon

Kellogg School of Management, 2001 Sheridan Road Evanston , IL 60208 , g-allon@kellogg.northwestern.edu

Achal Bassamboo

Kellogg School of Management, 2001 Sheridan Road Evanston , IL 60208 , a-bassamboo@kellogg.northwestern.edu

March 9, 2010

Many service providers use delay announcements to inform customers of anticipated delays. However, this information is usually not provided immediately, but rather after a short period of time (spent either waiting or occupied by the system). The focus of this paper is on the impact of this postponement on the ability of the firm to influence customer behavior by communicating non-verifiable congestion information to its customers as well as on the profits and utilities for the firm and the customers, respectively. We show that this postponement can actually help the firm create credibility and augment the resulting equilibrium. However, in other settings this delay can also detract from the resulting equilibrium. Further, we show that whenever credibility is created it improves not only the profit for the firm, but also the customers' overall utility under certain settings.

1. Introduction

In recent times, most service systems provide some form of delay-related information to their customers. In these systems, delay announcements provide prospective customers with information which contains an estimate of their waiting time if they decide to enter the system or informs them of the congestion level the system is currently experiencing. These announcements aim to improve both the customer service experience and the system performance. Most service organizations provide such information only after some delay has been experienced by the customer. The goal of this paper is to study this feature and its impact on the performance of the system, and the ability of firms to credibly communicate delay information to their customers.

In practice, firms use various types of messages, some being as precise as the expected waiting time in the queue or the number of customers ahead of you, whereas others are as vague as the statement “the system is experiencing long waits.” These messages signal to the customers the congestion the system is experiencing. In these cases, even after the firm categorizes the needs of the customer, it delays providing waiting time information. These delays can be inserted using many mechanisms. For example, call centers

employ Interactive Voice Response (IVR) to provide customers with current promotions information or the available rebates: the IRS uses a recorded message to inform customers of special dates and events such as rebates, and TIAA-CREF to inform customers of new products. The IVR may also be used to provide information about other channels which the customers can use: banks such as Washington Mutual and many airlines use the IVR to divert customers to their websites by informing them of this option.

In service systems, there are various methods to modulate demand. The key objective of these methods is to turn away customers (either voluntarily or involuntarily) when the system is experiencing high congestion. Admission control by the use of busy signals, or by immediately diverting customers to leave a voice message are examples of involuntary demand modulation. With the recognition that such involuntary admission control carries significant goodwill losses to the firm, many firms switched to voluntary methods where the customer is provided with information on the congestion of the system, leaving him the decision whether to stay or balk. Previous models of delay announcements have studied settings where information is announced immediately as the customers arrive and informs them of the firm's best estimate of the anticipated waiting time. All of these models (with exception of Allon et al. (2007)) assume that the customer treats this information as credible. Based on this information the customer then computes his expected utility and makes the decision whether to join or balk. As described above, many firms provide waiting time information after a delay, and thus postpone the demand modulation. In this paper, we study the impact of this postponement on the ability of the firms to provide unverifiable, non-committal real time information to its customers.

We treat the announcements made by the system manager as "cheap talk," i.e., pre-play communication that carries no cost. Cheap talk consists of costless¹, non-binding, non-verifiable messages that may affect the customer's beliefs. It is important to note that while providing the information does not *directly* affect the payoffs, it has an indirect implication through the customer's reaction and the equilibrium outcomes. The information has no impact on the payoffs of the different players per se i.e., the payoffs of both sides

¹ We assume that the cost associated with conveying the message is negligible. In most practical service organizations, while the provider needs to incur fixed costs, for example, by investing in a more sophisticated IT infrastructure to learn the state of the system, the marginal cost of providing the information to the customer is insignificant. There is a voluminous literature starting with Spence (1973) dealing with models where signaling is not costless, and the mere fact that players are willing to incur a cost provides a signal.

depend only on the actions taken by the customers and the queueing dynamics. This, in turn, means that if a customer does not follow the recommendation made by the firm, he is not penalized, nor is he rewarded when he follows it. However, as it will be shown, the announcements do have an impact on the service provider's profits and the customers' utility, in equilibrium.

Research questions and our model. In this paper, we study the role of delaying the delay announcement in enabling the delay announcements to impact customer behavior. Towards this objective, we consider two systems: The first system is one where the customers are provided the announcement immediately on arrival. (We refer to this setting as the base model.) In the second system, we consider a model in which the customer is first being delayed and then provided the announcement. In both models, operationally, the interaction following the announcement is exactly the same. In particular, we study the problem from the point of view of a firm that currently uses the base model, and considers adding the first stage, a common practice in the industry. Thus, it is important to understand the implication of postponing the information provision on the equilibrium emerging in the game as well as the firm's profits and the customers' experience.

We are interested in the impact of delaying the delay announcement on the equilibrium emerging in the game.

Before turning to the game theoretic analysis, the more basic operational question is whether such a postponement impacts the optimal admission control problem the firm solves when trying to determine whether to "recommend" the next customer to join or balk.

To analyze the game played between the firm and the customer in both settings (when the announcement is made upfront versus the case where announcement is delayed), one needs first to analyze these systems assuming the firm can dictate the decision to the customers. We refer to these problems as full control and full access control for the two settings, respectively. We begin by analyzing the full control problem, followed by the analysis of the game played between the firm and its customers when the information is provided immediately upon the customers' arrival. We then study the impact of postponement when the firm has full access control. Using these results, we study the impact of postponement on the game played between the customers and the firm when the information provided is treated as non-verifiable and non-credible. In particular, we compare the set of possible equilibria with and without postponement.

Our main contribution and results are as follows:

1. We characterize the optimal policy for the full access control problem. It is shown that for every number of customers in the first stage of the service process (i.e. the IVR), there exists a threshold on the number of customers in the second stage of the process above which the firm prefers rejecting the customer. This is referred to as a switching curve. It is important to note that since the transition rates in the first part of the service are not bounded, one cannot employ uniformization arguments to compute the control directly. However, using a bounding argument (which to the best of our knowledge is novel) with systems where uniformization can be employed we show that the optimal policy can be characterized. This technique may be employed in other Markov Decision Process analyses as well.

2. We characterize the set of possible equilibria in the delayed cheap talk game. Specifically, we provide conditions under which an equilibrium with influential cheap talk exists. We show that for such an equilibrium to exist the ratio between the customer's value of the service to his cost of waiting has to be within a band defined by two thresholds. A similar result applies to the non-delayed cheap talk model where the conditions for existence of an equilibrium with influential cheap talk can be described using two different threshold levels. We also show that for the delayed cheap talk game, as the value obtained from the IVR decreases, the width of the band may diminish.

3. We systematically compare the set of equilibria arising in the delayed cheap talk model with the one arising in the non-delayed game, and assess the value of postponement. We show that there are instances where the firm can create credibility and impact customers' behavior by delaying the delay announcements. However, it might also *lose* its credibility and its ability to impact customers' behavior due to this delay. Under specific conditions on the value obtained from the IVR, we show that both the firm and the customers always prefer an equilibrium with influential cheap talk over equilibria with non-influential cheap talk.

4. We also discuss the case where the IVR is an essential part of the service process and thus a firm might consider providing the information even before this stage. We show that, from the firm's point of view, delaying the information provision is beneficial regardless of whether the IVR stage is essential or not.

5. We show that under certain conditions our findings can be extended to the case where the customers evaluate their option of staying in the system versus leaving the system continuously. Thus, customers have the option to abandon the system. We show that for the equilibrium prescribed in the paper the customer would never exercise this option. In this sense these equilibria are abandonment-proof.

Organization of the paper. In the next section, we discuss the relevant literature. Section 3 describes the base-model where the information is provided upfront. Section 4 describes and analyzes the full access control problem where the customer is being delayed. Section 5 describes and analyzes the delayed cheap talk game. Section 6 contrasts the equilibrium strategies as well the outcomes. This section also provides numerical study. Section 7 discusses extensions of the models. Section 8 provides the conclusion to the paper. The proofs of all the theorems appear in the body of the paper, following the statement of the theorem. The proofs of propositions are provided in the Appendix.

2. Literature Review

Delay announcement models. There is a growing interest in models studying the impact of delay announcements on the system performance when the queue is invisible. One of the first papers that discusses this issue is Hassin (1986) which studies the problem of a price-setting, revenue-maximizing service provider that has the option to reveal the queue length to arriving customers, but may choose not to disclose this information. It is shown that it may be – but not always – socially optimal to prevent suppression of information, and that it is never optimal to encourage suppression when the revenue maximizer prefers to reveal the queue length. Whitt (1999), assuming that information provision reduces the likelihood of balking, shows how services can be improved by informing customers about anticipated delays. Armony and Maglaras (2004a) and Armony and Maglaras (2004b) study systems where the service manager provides the customers an estimate of the delay, based on the state of the system upon their arrival.

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 told upon arrival of either the delay of the last customer to enter service, or an appropriate average delay. The authors show that within the fluid-model framework, under certain conditions, the actual delay coincides with the

announced delay. Guo and Zipkin (2007) studies a model in which customers are provided with information and make decisions based on their expected waiting times, conditional on the provided information. The authors consider three settings where (a) no information is provided, (b) queue length information is provided, or (c) accurate waiting time information is provided. They show that accurate delay information may improve or hurt the system performance. Motivated by this type of delay announcement, Ibrahim and Whitt (2008), explores the performance of different real time delay estimators based on recent delay experience by customers. Jouini et al. (2007) studies a model where customers react by balking upon hearing the delay announcement, and may subsequently renege if the realized waiting time exceeds the delay that has originally been announced to them. The balking and renegeing from such a system are a function of the delay announcement precision. The authors, analytically, characterize the performance measures for this model, and using these within a numerical study, explore when informing customers about delays is beneficial, and what the optimal precision should be in these announcements. For empirical evidence of how customers are impacted by delay announcements, see Hui and Tse (1996) and Kumar et al. (1997) and for empirical study of the impact of service variability, see Kumar and Krishnamurthy (2008).

Admission control in tandem queues/network. Our paper is also related to the literature on admission (access) control in multi-stage queueing systems. Ghoneim and Stidham (1985) studies the problem of admission control in a two queue tandem network with input to each queue where the system manager can accept or reject an arriving customer. Using a dynamic programming approach the authors show monotonicity properties which are used to derive structural properties of the optimal control. Ku and Jordan (2002) studies the admission control problem in a two-stage-loss-system. The authors prove that the optimal admission control policy is given by a set of thresholds. In these papers, the control policy must reject an (internal or external) arriving customers if the station does not have an idle server. The optimal policy is shown to be the one where the internal customers are never rejected if one of the servers is idle. This is due to the fact that the cost of blocking an accepted customer is exorbitantly high, compared to blocking a new customer (Ku and Jordan (2003) extends this for multiple stations in tandem). This is in contrast to our model, where the customers themselves terminate their call or request for service based on *their* assessment of the state of the system.

Classical Cheap Talk. The framework used in this paper echoes the classical cheap talk model proposed in Crawford and Sobel (1982). Crawford and Sobel (1982) introduced the Sender-Receiver cheap talk game to study strategic information transmission. In their model, the sender has private information, and the receiver takes payoff-relevant actions. The distribution of the sender's private information is fixed exogenously and does not depend on the equilibria of the game. This is in contrast to our endogenous cheap talk setting, where 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 receivers (customers) whose actions have externalities on other receivers; and second, the stochasticity of the state-of-the-world (i.e., 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 depend on the equilibrium strategies of both the firm and the customers. In particular, in our model, the customers' actions are payoff-relevant as well as system-dynamic-relevant. As we shall see, the multiplicity of receivers with externalities as well as the endogenization of the uncertainty impact both the nature of the communication as well as 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.

Delay announcements as cheap talk. Allon, Bassamboo, and Gurvich (2007) appears to be the first paper in the operations management literature to consider a model in which a firm communicates unverifiable real time dynamic delay information to its customers. Our paper is closely related to this paper in terms of the underlying framework. Both papers focus on analyzing the problem of information communication in an operational setting by considering a model in which both the firm and the customers act strategically: the firm in choosing its announcements, and the customers in interpreting this information and in making the decision. While the Allon, Bassamboo, and Gurvich (2007) focuses on developing a framework for cheap talk in service settings, the focus of the current paper is to understand the impact of delaying the delay announcement when the information provision is strategic both for the customers and the firm.

3. Base Model: Benchmark

In this section, we first develop a model where the service process contains only a single stage and the service provider announces the delay-related information immediately to an arriving customer. This case will serve as a benchmark when we shall study the scenario of delaying the information transmission.

For this base model, we consider a service provider modeled as an M/M/N system. Customers arrive to the system according to a Poisson process with rate λ . Service times are exponentially distributed with mean $1/\mu$. We assume that $\lambda < N\mu$. We assume that all customers are ex-ante symmetric: customers obtain a value R if they are served, and incur a waiting cost that is proportional to the time spent in the system, with a unit waiting cost of c . Thus, a customer arriving to the system obtains the following utility:

$$U(y) = \begin{cases} R - cw & \text{if } y = \text{“join,”} \\ 0 & \text{if } y = \text{“balk,”} \end{cases} \quad (1)$$

where y is the decision made by this customer and w denotes its waiting time in the system. When a customer arrives, the system manager has private information regarding the number of customers currently waiting in queue, denoted by the random variable Q . Its distribution will depend on the equilibrium strategies of both the provider and the customers. We assume that the customer decides whether to join or not based on the information he can infer from the system manager regarding the current state of the system in order to maximize its expected utility. We denote the information by I . The best estimate of the customers waiting time in the queue is given by $\mathbb{E}[w|I]$, i.e., conditioned on the information the customer is provided he computes the expected waiting time. Thus, the customer will join, if $R \geq c\mathbb{E}(w|I)$. The system manager obtains revenue of v per customer served, and incurs a holding cost h per unit of time per customer. Note that this waiting cost constitutes, among other factors, the cost associated with loss of goodwill, the actual cost of holding the customer and in some settings the opportunity cost associated with the customer not being able to generate revenues. For example, Disney’s theme parks incur two costs due to waiting: the opportunity cost of having a customer kept in line without the ability to spend money and the wages of the entertainment staff that is in charge of alleviating the pain of waiting. The firm’s profits are then given by the following expression:

$$\mathbb{E} \left[\int_0^\infty e^{-\alpha t} v dD(t) - \int_0^\infty e^{-\alpha t} h Q(t) dt \right],$$

where $D(t)$ is the departure process from the system, and α is the discount factor.

Next, we formally define the game between the service provider and the customers. The equilibrium concept we employ is one of Markov Perfect Bayesian Nash Equilibrium (MPBNE), which, in this case, is simply a Nash equilibrium in the decision rules that relate agents' actions to their information and to the situation in which they find themselves, while allowing for actions to depend only on payoff-relevant histories. Let \mathcal{M} denote the Borel set which is comprised of feasible signals that the firm can use. We represent the signaling rule by a function $g : \mathbb{Z} \mapsto \mathcal{M}$, where $g(q) = m$ if the firm uses the signal m when the number of customers in the system is q . Let $y : \mathcal{M} \mapsto \{0, 1\}$ denote the strategy of the customer, where $y(m)$ is the probability that a customer joins when the firm signals m . Consequently, we interpret $y(m) = 1$ as a "join" decision and $y(m) = 0$ as a "balk" decision and we will use this alternative terminology interchangeably. Note that the above signaling and action rules restrict attention to pure strategies. Noting that the strategies are only state dependent, the resulting system dynamics can be modeled using a birth-death process. Further, since the arrival rate λ is less than the service capacity μN the resulting birth-death process, under any (y, g) , has a unique steady-state distribution. We denote this using $p_q(y, g)$ where $p_q(y, g)$ is the steady-state probability of having q customers in the system when the firm employs signaling rule g and the customers follow the decision rule y .

Definition 3.1 (Markov Perfect Bayesian Nash Equilibrium) *We say that the signaling rule $g(q)$ and the action rule $y(m)$ constitute a Markov Perfect Bayesian Nash Equilibrium (MPBNE), if they satisfy the following conditions:*

1. *For each $m \in \mathcal{M}$, we have*

$$y(m) = \begin{cases} 1 & \text{if } \frac{\sum_{\{q: g(q)=m\}} \left[\frac{R-cq-N+1}{N\mu} \right] p_q(y, g)}{\sum_{\{q: g(q)=m\}} p_q(y, g)} \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

2. *There exist constants J_0, J_1, \dots , that solve the following set of equations:*

$$\begin{aligned} J_q &= \max_{m \in \mathcal{M}} \left\{ \frac{hq}{\lambda + \mu + \alpha} + \frac{\mu}{\lambda + \mu + \alpha} [(J_{q-1} + v)\mathbb{I}(q > 0) + J_0\mathbb{I}(q = 0)] + \frac{\lambda}{\lambda + \mu + \alpha} (J_q(1 - y(m)) + J_{q+1}y(m)) \right\} \\ &= \left\{ \frac{hq}{\lambda + \mu} + \frac{\mu}{\lambda + \mu} (J_{q-1} + v) + \frac{\lambda}{\lambda + \mu} (J_q(1 - y(g(q))) + J_{q+1}y(g(q))) \right\} \end{aligned} \quad (2)$$

In the above definition of MPBNE, the first condition represents the customer's best response to the firm's announcement strategy, using the Bayesian rule. The second condition states that the composite function $y \circ g$ solves the firm's MDP which determines the firm's best response to the customer's strategy.

One of the goals of this paper is to identify the conditions under which a firm can credibly communicate unverifiable information. Our litmus test for such credibility will be the existence, or lack thereof, of an equilibrium with influential cheap talk. When such an equilibrium exists, it means that the firm can induce, by virtue of using at least two distinct messages, two distinct actions. An equilibrium with influential cheap talk is formally defined as follows:

Definition 3.2 *We say that an equilibrium (y, g) has influential cheap talk if there exists two signals m_i, m_j where $i \neq j$ such that $y(m_i) \neq y(m_j)$, $\sum_{\{q:g(q)=m_i\}} p_q(y, g) > 0$ and $\sum_{\{q:g(q)=m_j\}} p_q(y, g) > 0$.*

Following the arguments used in Propositions 3.1 and 3.2 of Allon et al. (2007) and noting the fact that the underlying state-space dynamics form a birth-death chain in equilibrium, we can reduce the strategy space to a threshold. Further, we can show that when an equilibrium with influential cheap talk exists, any influential equilibrium is equivalent to an equilibrium where the firm uses only two signals: one to signal that the system has low congestion and one that signals high congestion, and thus advises the customer to balk.

As in Allon et al. (2007), in order to characterize the possible equilibria (which are threshold induced) under different settings, we shall first characterize two important threshold levels: the first, q^* , denotes the threshold value above which a customer *will not join* if he has **full information** of the state of the system, and below which he *will join*. The second threshold level, \hat{q} , is motivated by the service provider's point of view, and denotes the threshold level below which the service provider would like the customers to join, and above which she would like them to balk, if she had **full control** over their actions.

Full information. We will define q^* to be the threshold value above which the customer will not obtain positive utility, in expectation, given full queue length information. It is easy to see that

$$q^* = \left\lfloor \frac{RN\mu}{c} \right\rfloor - 1 + N, \quad (3)$$

where $\lfloor \cdot \rfloor$ is the floor function; i.e., q^* is the largest integer not exceeding $RN\mu/c - 1 + N$. Note that this threshold pertains to the marginal customer who decides to balk.

Full control. From the service provider's point of view, deciding on a threshold level amounts to deciding what should be the finite waiting space in an $M/M/N/k$ queueing system where k denotes the maximum number of customers present in the system at any time. For each value of k , let D^k denote the departure process and Q^k denote the queue length process. Thus, if the firm decides the threshold level to be k for the number of customers in the system, its expected profit is given by

$$\Pi(k) = \mathbb{E} \left[\int_0^\infty e^{-\alpha t} v dD^k(t) - \int_0^\infty e^{-\alpha t} h Q^k(t) dt \right]. \quad (4)$$

Let \hat{q} denote the optimal waiting space, i.e., it solves the following full control optimization problem $\hat{q} \in \arg \max_k \Pi(k)$. (The existence of \hat{q} can be shown using Knudsen (1972).)

Next note that the customer's expected utility if he finds q customers in the system upon arrival and decides to join the queue is given by $R - c(q + 1 - N)/N\mu$. Thus, if the firm wants to turn away customers when there are \hat{q} customers in the system, it must be the case that $R - c\frac{\hat{q}-N+1}{N\mu} < 0$. For future purposes, we define $\bar{\eta}^B = \hat{q} - N + 1$. Further, suppose that the firm provides one signal for all states when the number of customers is strictly less than \hat{q} but strictly greater than N . Denoting the expected number of customers in system when the firm provides this signal by \tilde{q} , we obtain that the customers will join the system upon receiving that signal if and only if $R - c\frac{\tilde{q}-N+1}{N\mu} \geq 0$. We can then define $\underline{\eta}^B = \tilde{q} - N + 1$. Note that $\underline{\eta}^B \leq \bar{\eta}^B$.

Based on the above analysis we obtain the following result that characterizes the influential equilibria of the above cheap talk game.

Proposition 3.1 *We have that the above cheap talk game has a pure strategy MPBNE with influential cheap talk if and only if*

$$\underline{\eta}^B \leq \frac{RN\mu}{c} \leq \bar{\eta}^B. \quad (5)$$

Further, the pair $(g(\cdot), y(\cdot))$ defined by

$$g(q) = \begin{cases} m_1 & q \leq \hat{q}, \\ m_0 & \text{otherwise.} \end{cases}, \quad y(m) = \begin{cases} 1 & m = m_1, \\ 0 & \text{otherwise.} \end{cases} \quad (6)$$

forms an equilibrium with influential cheap talk.

It is interesting to note that we can also state the above result in terms of \hat{q} and q^* as shown in Allon et al. (2007), however the above representation would aid us in understanding the impact of postponing the delay announcement.

To summarize the findings so far: we have identified a region in which a pure strategy MPBNE exists. Note that in this region, using the definition of q^* and $\bar{\eta}^B$ we must have $q^* \leq \hat{q}$. Thus for an equilibrium to exist, the firm's and the customers' incentives are either perfectly aligned or the customers are mildly impatient. When an equilibrium with influential cheap talk does not exist, Allon et al. (2007) shows that a babbling equilibrium in which all signals result in the same behavior may exist in pure strategies.

4. Model with delayed announcement

In many service systems firms provide information only after a certain period of time. Most firms use the time in which the customer is waiting (prior to providing the information) to inform him of current promotions (for example, Dominos Pizza announces the current deals) or specific events, thus both the firm and the customer are obtaining some value, while incurring some waiting and holding costs during this phase. The key questions we aim to answer are (i) how does the ability to postpone the information provision impact the emerging equilibria? (ii) Under what circumstances can a firm sustain equilibria with influential cheap talk (and achieve its first best profit)?

Note that without postponement a firm may not admit a customer even if he brings positive revenue to the system. The firm needs to solve a non-myopic admission control policy taking into account the externality the customer imposes on the other customers and hence the profit of the firm. However, since the firm needs to make the decision immediately upon arrival of the customer, it uses the *expected* externality to make this choice. To understand the role of delay, imagine a clairvoyant firm making an admission decision. In this case, it would know the exact arrival process and hence the true externality of this customer. Hence, these decisions would undoubtedly perform better. If the firm can delay the admission decision, it can get a better estimate of the externality imposed by a given customer on the system. In this section, we explore a system in which the customers are provided the announcement after some delay. The question of how this delay impacts the ability to influence customers' behavior remains unanswered and is the focus of this section.

We shall next define the model for the delayed announcements. We consider a system where the customers arrive according to a Poisson process. The service process consists of two stages: the first stage, which is not necessary for the service, but can create value² for the two sides, essentially does not require any resources. The second stage which generates the value for the firm and the customer requires human resources for its completion. In a call center setting, the first stage is usually done through an Interactive Voice Response (IVR) which is usually used to provide the customers with some background information, as well as to gather information from the customer in an efficient manner. We shall assume that the IVR has sufficient capacity so no customer has to wait for it, which we model as an infinite server queue. In this setting, the customer will be provided information only upon completion of the IVR stage. We thus treat the IVR as the delay mechanism and use it to study the impact of delaying the delay announcement.

The sequence of events and the service process model. The interaction between the customers and the firm can be described as follows: the customers arrive to the system according to a Poisson process. They first are faced with the IVR system, which we model as an $M/M/\infty$ system. The rate at which the jobs get processed is exponential with rate μ_{IVR} . (See Feigin (2006) for motivation for stochastic time at the IVR.) After the customers complete the interaction with the IVR, the system manager provides some message with regards to the waiting in the actual system (the agent based system). At this point the customer can decide to join or renege the system. If he decides to join, he enters the queue for the Agent-Based Service (ABS) which we model as a multi server queue with N servers. We shall assume that the customer never reneges the system once he enters the agent-based-service (ABS). (We shall investigate the system where the customer has an option to renege after joining in Section 7.1.)

Utility of the customer. Consider an arriving customer to the system. Once the customer is processed by the IVR system, he would receive a signal regarding the system congestion, based on which he would decide to stay or balk. Depending on this decision and the actual waiting time, his ex-post utility is described as follows: the customer obtains a utility of R if served and incurs a cost of c per unit of time waiting. Thus, the utility of the customer equals $v_{IVR} - cw_{IVR}$ if he balks after the announcement, and $R_{IVR} + R - c(w_{IVR} +$

² For example, consider customers calling to purchase an airline ticket. In many instances, the airline provides information regarding new security restrictions and guidelines. Clearly, this information is not essential for the ticket purchasing process. However, this information is beneficial for the customer and the airline, and thus creates value for both sides.

w) where w_{IVR} is the waiting time in the IVR and w is the actual waiting time for the agent based service.

Thus, a customer arriving to the system obtains the following utility:

$$U^D(y) = \begin{cases} (R_{IVR} + R) - c(w_{IVR} + w) & \text{if } y = \text{“join,”} \\ (R_{IVR} - cw_{IVR}) & \text{if } y = \text{“balk,”} \end{cases} \quad (7)$$

where y is the decision made by this customer. Based on this, it is clear that customers balk only if $R \geq c\mathbb{E}[W|I]$, where $\mathbb{E}[W|I]$ is the expected waiting time to the ABS in equilibrium given the information provided by the system manager is I . (Note that the expected waiting time of a customer to the ABS depends on other customers' actions, thus one needs to define an equilibrium among the customers. We shall define this equilibrium concept formally and also take into account this dependence in Section 5.)

Profit of the firm. The firm receives v_{IVR} from every customer who arrives to the system. Further, for the customers who get served, the firm receives a value v upon each service completion. At the same time, the firm also incurs a holding cost of h per customer per unit of time for any customer in the system (irrespective of the customer being in the IVR or the ABS³). Let $Q_{IVR}(t)$ and $Q(t)$ denote the number of customers in the IVR and the ABS at time t , respectively. Let $D(t)$ be the counting process corresponding to the departure from the ABS. Then the firm's profit function is given by

$$\mathbb{E} \left[\int_0^\infty e^{-\alpha t} [vdD(t) - h(Q_{IVR}(t) + Q(t))dt] \right] + \frac{\lambda v_{IVR}}{\alpha},$$

The first term $vdD(t)$ corresponds to the fact that the firm obtains a value of v for each service completion (which is equivalent to departure from the agent-based service). The second term $h(Q_{IVR}(t) + Q(t))$ corresponds to the holding cost incurred by the firm, proportional to the number of customers waiting at each point in time. The last term denotes the revenue generated by the IVR system.

4.1. Full Information Solution.

Suppose that at the decision instance, (which corresponds to the moment at which the IVR completes the processing of the customer) the customer has full information with regards to the system status. The optimal decision based on this information is easy to characterize. The customer arriving to the ABS decides to balk

³ The analysis followed in the paper can be easily extended to the setting where the firm incurs different holding cost for customers in the IVR and the ABS

the system after he is processed at the IVR only if $c(Q(t) + 1 - N)/N\mu > R$ where $Q(t)$ is the number of customers in the ABS. At the decision instance, a rational customer, who is trying to maximize his overall utility, will find that the experience at IVR is already realized and (for the purpose of making the best decision) is thus not a function of the action. Thus, his optimal decision is driven only by the utility he obtains from the ABS. Further, this decision (which is based on $Q(t)$) is the dominant strategy for each deciding customer, as the customer is always better off following it, regardless of the decisions made by other customers. This is due to the fact that all customers arriving to the ABS, following a given customer, cannot impact his waiting time, as the system manager is following a first-come-first-served service rule. Thus the decision is entirely based on the number of customers currently in the ABS when the deciding customer arrives to the ABS. Further, the threshold on the number of customers in ABS above which the customer decides to balk the system is identical to the one used by the customer in the base model where there was no delay in providing information.

4.2. Full Access Control Solution

Next, we will analyze the problem from the firm's point of view: suppose the firm could make decisions for the customers who have been processed by the IVR whether to join the ABS or not, once their service is completed. We shall refer to such a system as one with full control. In this setting, the state descriptor is a vector $Q^S(t) \equiv (Q_{IVR}(t), Q(t)) \in \mathbb{Z}^2$, where \mathbb{Z} denotes the set of whole numbers $\{0, 1, \dots\}$. Here, $Q_{IVR}(t)$ denotes the number of customers in IVR at time t , and $Q(t)$ denotes the number of customers in the ABS at time t . When the customer completes his service at the IVR, the system manager makes the accept/reject decision (based on which the customer would be admitted to the ABS or turned away) based on the state of the system $(Q_{IVR}(t), Q(t))$. Note that in doing so, the system manager is not only taking into account the expected wait this customer would experience (which is simply a function of Q) but also the "externalities" he imposes on other customers (which depend on $Q_{IVR}(t)$ and future arrivals), and the ability of the firm to generate profits from them. We next show that the optimal access policy is threshold-based.

Theorem 4.1 *There exists a threshold function $\eta^*(\cdot)$, such that it is optimal for the firm to accept a customer that completed service at the IVR, if and only if $Q \leq \eta^*(Q_{IVR}(t))$ and "turn him" away otherwise.*

Proof:

To prove the result, we shall use the Markov Decision Process theory. However, note that it is not possible to employ the concept of uniformization directly on the system as the service rate in IVR system grows without bound. To this end, we shall consider a sequence of systems index by κ . The κ^{th} system is identical to the one defined before with the modification that the IVR has κ servers and there is no waiting in the IVR, i.e., IVR is a pure loss system. We will make another transformation that will be useful for analyzing this system, we assume that the customer pays v immediately upon entering the queue for the ABS, and the holding cost of a customer in IVR is $h + \alpha v$. It is easy to see that the above is equivalent to the setting where the holding cost in ABS is h , and the customer pays v upon service completion. For this system we can use the uniformization approach to obtain the following optimality equations:

$$\begin{aligned} V^\kappa(q_{IVR}, q) &= \frac{1}{\alpha + \kappa\mu_{IVR} + N\mu + \lambda} [-hq_{IVR} - h(q - N)^+ - \alpha vq \\ &\quad + \lambda V^\kappa(q_{IVR} + \mathbb{I}_{q_{IVR} < \kappa}, q) + \mu(\min\{q, N\}V_{n-1}^\kappa(q_{IVR}, q - 1) + (N - q)^+ V_{n-1}^N(q_{IVR}, q)) + \\ &\quad + \mu_{IVR} \min\{q_{IVR}, \kappa\} \max\{V^\kappa(q_{IVR} - 1, q + 1) + v, V^\kappa(q_{IVR} - 1, q)\} + (\kappa - q_{IVR})\mu_{IVR} V^\kappa(q_{IVR}, q)] \end{aligned}$$

Here x^+ is define as $\max\{0, x\}$.

Next, we define a map \mathcal{L} on real valued functions on \mathbb{Z}^2 , and a sequence V_n^N for $n = 1, 2, \dots$, where $V_0^\kappa \equiv 0$ and $V_n^\kappa = \mathcal{L}V_{n-1}^\kappa$ is given by

$$\begin{aligned} V_n^\kappa(q_{IVR}, q) &= \frac{1}{\alpha + \kappa\mu_{IVR} + N\mu + \lambda} [-hq_{IVR} - h(q - N)^+ - \alpha vq + \lambda V_{n-1}^\kappa(q_{IVR} + \mathbb{I}_{q_{IVR} < \kappa}, q) + \\ &\quad + \mu(\min\{q, N\}V_{n-1}^\kappa(q_{IVR}, q - 1) + (N - q)^+ V_{n-1}^N(q_{IVR}, q)) + \\ &\quad + \mu_{IVR} \min\{q_{IVR}, \kappa\} \max\{V_{n-1}^\kappa(q_{IVR} - 1, q + 1) + v, V_{n-1}^\kappa(q_{IVR} - 1, q)\} \\ &\quad + (\kappa - q_{IVR})\mu_{IVR} V_{n-1}^\kappa(q_{IVR}, q)]. \end{aligned}$$

It is clear that V^κ is a fixed point of this mapping. Using the theory of the semi-Markov decision process, we also have that $V_n^\kappa \rightarrow V^\kappa$ as $n \rightarrow \infty$. To show concavity of V^κ , we shall show that V_n^κ is concave and non-increasing in q . The proof is by induction. The result holds for $n = 0$ by definition as $V_0^\kappa \equiv 0$. Assume that V_{n-1}^κ is concave and non-increasing in q we shall show that V_n^κ is concave and non-increasing in q . First, note that $-hq_{IVR} - h(q - N)^+ - \alpha vq$ is concave, and $\lambda V_{n-1}^\kappa(q_{IVR} + \mathbb{I}_{q_{IVR} < \kappa}, q)$ and $(\kappa -$

$q_{IVR})\mu_{IVR}V_{n-1}^\kappa(q_{IVR}, q)$ are both concave in q , as V_{n-1}^κ is concave in q . Second, consider $f_1(q_{IVR}, q) \equiv (\min(q, N)V_{n-1}^\kappa(q_{IVR}, q-1) + (N-q)^+V_{n-1}^\kappa(q_{IVR}, q))$. Using the fact that V_{n-1}^κ is concave and decreasing in q and appealing to Lemma 3.1 Koole (1998), we obtain that $f_1(q_{IVR}, q)$ is decreasing and concave in q . Lastly, define $f_2(q_{IVR}, q) \equiv \max\{V_{n-1}^\kappa(q_{IVR}-1, q+1) + v, V_{n-1}^\kappa(q_{IVR}-1, q)\}$. Fix q_{IVR} . Noting that $V_{n-1}^\kappa(q_{IVR}-1, q)$ is concave decreasing in q , there exists a q^* such that $V_{n-1}^\kappa(q_{IVR}-1, q) - V_{n-1}^\kappa(q_{IVR}, q+1) \leq v$ if $0 \leq q \leq q^*$, and $V_{n-1}^\kappa(q_{IVR}-1, q) - V_{n-1}^\kappa(q_{IVR}, q+1) > v$ if $q > q^*$. Thus, we have that

$$f_2(q_{IVR}, q) = \begin{cases} V_{n-1}^\kappa(q_{IVR}-1, q+1) + v & \text{if } q \leq q^* \\ V_{n-1}^\kappa(q_{IVR}-1, q) & \text{otherwise.} \end{cases} \quad (8)$$

We need to show that $f_2(q_{IVR}, q)$ is concave and non increasing in q . Since $V_{n-1}^\kappa(q_{IVR}-1, q)$ is concave and non-increasing, it suffices to show the following:

$$f_2(q_{IVR}, q^*) \geq f_2(q_{IVR}, q^*+1) \quad (9)$$

$$f_2(q_{IVR}, q^*-1) - f_2(q_{IVR}, q^*) \leq f_2(q_{IVR}, q^*) - f_2(q_{IVR}, q^*+1) \quad (10)$$

The first follows by noting that $v \geq 0$ and the representation of f_2 in (8). For the second equality note that using (8) the left hand side equals $V_{n-1}^\kappa(q_{IVR}-1, q^*) - V_{n-1}^\kappa(q_{IVR}-1, q^*+1)$ and the right hand side equals v . The second equality then follows by the definition of q^* .

Thus, we get that $V_n^\kappa(q_{IVR}, q)$ is concave and non-increasing in q . Thus, we establish the fact that $V^\kappa(q_{IVR}, q)$ is concave.

Next we will show that for large κ , the system performance obtained, when modelling the IVR as $M/M/\kappa/\kappa$ system is close to the one obtained when modeling it as an $M/M/\infty$ system. We define two random variables: $Q_{M/M/\kappa/\kappa}$ which represents the number of customers in an $M/M/\kappa/\kappa$ queue in steady state with arrival rate λ and service μ_{IVR} ; and $Q_{M/M/\infty}$ which represents the number of customers in an $M/M/\infty$ queue in steady state with arrival rate λ and service μ_{IVR} . Let $V(q_{IVR}, q)$ be the optimal profit for the firm when the IVR has infinite capacity starting from the state (q_{IVR}, q) . Then, it is easy to see that

$$V^N(q_{IVR}, q) - \lambda\mathbb{P}(Q_{M/M/\kappa/\kappa} = N)h\mathbb{E}[w] \leq V(q_{IVR}, q).$$

Further, we know that $\mathbb{E}[w] = 1/\mu_{IVR}$, and $\mathbb{P}(Q_{M/M/\kappa/\kappa} = \kappa) \leq \mathbb{P}(Q_{M/M/\infty} \geq \kappa)$. Also, we can show

$$V(q_{IVR}, q) \leq V^\kappa(q_{IVR}, q) + \lambda v\mathbb{P}(Q_{M/M/\kappa/\kappa} = \kappa) \leq V^\kappa(q_{IVR}, q) + \lambda v\mathbb{P}(Q_{M/M/\infty} \geq \kappa).$$

Noting that the number in system for an $M/M/\infty$ system is Poisson distributed with mean $\lambda\mu_{IVR}$, there exists $\beta > 0$ such that for N large

$$\mathbb{P}(Q_{M/M/\infty} \geq N) \leq e^{-\beta\kappa}.$$

Thus we have for N large, there exist a finite K such that

$$\|V^\kappa - V\| \leq Ke^{-\beta\kappa}.$$

Thus, one gets $V^\kappa \rightarrow V$ as $\kappa \rightarrow \infty$. Thus, concavity of V^κ in q results in concavity of V in q . Hence, we obtain that there exists a threshold $\eta(q_{IVR})$ such that the firm would accept the customer into ABS from IVR if and only if $q \leq \eta(q_{IVR})$. Thus we have the optimal policy is a threshold policy. ■

The above result establishes the existence of a threshold-based access control policy: in order to maximize its long-run discounted profits, the firm should “accept” customers as long as the number of customers waiting for the agent based service is below a certain level. This level depends on the number of customers occupied by the IVR, such that, with q customers waiting for the ABS, a customer completing the IVR stage should be admitted to the ABS, only if $q < \eta^*(q_{IVR})$. For the rest of the paper, we will assume the following:

Assumption 4.1 *The threshold η^* , that solves the Full Access Control problem, is unique.*

The existence of a threshold amounts to showing that, fixing the number of customers occupied by the IVR, if the firm should reject a customer with q customers waiting for the ABS, it should do so for any number of customers waiting above q . The structure of the threshold function however is difficult to characterize and this stems from the fact that the service rate in the IVR is state-dependent, a similar issue is also raised in Ghoneim and Stidham (1985). While the existence of a threshold function is essential to derive the emerging equilibrium language in the next section, the exact structure or any monotonicity property of the threshold would not effect the results.

The intuition behind why the delay could be beneficial to the firm is as follows: During the time lag between the arrival and the announcement of the delay, more customers enter the system and hence the

firm may obtain a better estimate of the level of externalities the customer inflicts on other customers. The customer's externality on each other impacts their waiting time. Since the service provider also incurs the cost of waiting for the customers he cares for this externality. While in the single-stage model a firm has to decide whether to admit a customer based on his expected externality on other customers, in the two-stage model, the firm makes a more informed decision, and thus is able to achieve a higher profit, if able to implement the first-best solution. Next, we will characterize the circumstances under which the firm can sustain an equilibrium with influential cheap talk, while possibly achieving its first best profit.

It is worth noting that even though the output from the IVR is a Poisson process, it is not Poisson when we condition on the number of customers in the IVR. Thus, the firm uses the state of the system that consists of number of customers in IVR and ABS, the optimal control will (barring certain parameter choices) depend on both. Further, a threshold that is based only on the ABS is a feasible control and hence the firm's profit can only improve by using the information regarding the state of the system (that includes the number of customers in IVR).

5. The Delayed Cheap Talk Game

In the previous section we showed that if the firm has full control over the system in terms of which customers join the ABS, the firm would employ a threshold based control, i.e., based on the congestion in the ABS and IVR it would decide whether the customer who completes his service at the IVR gets transferred to the ABS or is removed/balks from the system. Further, in the settings where the customers were able to see the state of the system and make their own decisions they would completely ignore the congestion in the IVR and would use a fixed threshold policy based on the number of customers in the ABS. Thus, we see that the firm and its customers are not perfectly aligned. In reality, the customers do not have the information about the state of the system and the firm can use announcements regarding the state of the system to induce a desired customer behavior. However, it is crucial to note that since the information is unverifiable, the customers treat any information provided by the firm as a priori non-credible, unless the firm is able to gain credibility, in equilibrium.

The question we want to study is under what circumstances can the firm and the customers establish an equilibrium with influential cheap talk. In the base model, when the announcements are made immediately

upon the arrival of the customer, we have seen that equilibria with influential cheap talk exists only if condition (5) holds. We are interested in understanding how delaying the announcement can create (or destroy) credibility by the firm.

To study this formally, we shall begin by defining the delayed cheap talk game and the strategies for the firm and its customers. As before, \mathcal{M} denotes the Borel set which is comprised of feasible signals that the firm can use. Let $y : \mathcal{M} \mapsto \{0, 1\}$ represent the strategy of the customer who completed the service at IVR and is about to go to the ABS. Here, $y(m)$ takes value 1 or 0, if the customer joins the ABS or abandons the system, after completing his/her service at IVR and receiving a signal $m \in \mathcal{M}$, respectively. Let the space of feasible strategies for the customer be denoted by \mathcal{Y} . Let $g : \mathbb{Z}^2 \mapsto \mathcal{M}$ represent the strategy of the firm. Here $g(q_{IVR}, q)$ represents the announcement that the firm makes to the customer completing service at the IVR when the state of the system is (q_{IVR}, q) . Let the space of feasible strategies for the firm be denoted by \mathcal{G} . Note that the steady-state distribution of the number of customers in the ABS is determined by the customer's strategy y as well as the firm's strategy g . Let $p_{y,g}(q)$ represent the probability that in steady-state the number of customers in the ABS is q , if the firm follows strategy g and the customers follow strategy y . Further, let the firm's profit under the strategy pair y, g be written as $\Pi(y, g)$.

Definition 5.1 *We say that the pair $(y, g) \in \mathcal{Y} \times \mathcal{G}$ forms a Markov Perfect Bayesian Nash Equilibrium (MPBNE) in the delayed announcement game if and only if it satisfies the following two conditions:*

1. For all $m \in \mathcal{M}$,

$$y(m) \in \arg \max_{y \in \{0,1\}} y \left[\frac{\sum_{\{q:g(q)=m\}} [R - c \frac{q-N+1}{\mu N}] p_{y,g}(q)}{\sum_{\{q:g(q)=m\}} p_{y,g}(q)} \right].$$

2. y and g satisfy the following: $\mathbb{E}[U(y(g)) > 0]$.
3. Fixing y, g solves:

$$g \in \arg \max_{\tilde{g} \in \mathcal{G}} \Pi(y, \tilde{g}).$$

The above definition is closely related to the one defined for the base model. It also requires that the customers are not only making the appropriate choice based on the announcement at the beginning of the ABS but are also interested in the system only if their net utility from the entire experience (IVR and

ABS) yields a positive utility. The second condition was not required in the base model, as Condition 1 of Definition 3.1 for the customers ensured that they obtain positive utility from the system. It is important to note that we do not assume that the net utility a customer obtains from the IVR is positive. That is, $R_{IVR} - c/\mu_{IVR}$ can be negative.

5.1. Existence of Pure Strategy Equilibria in the Cheap Talk Game

In this section, we characterize conditions under which pure strategy equilibria with influential cheap talk exist. We shall show that the queuing dynamics observed under any MPBNE with influential cheap talk (if it exists) corresponds to the one where the firm achieves its first best, i.e., the firm has full control. Based on this observation, consider the system where the firm implements the Full Access Control solution. Let the steady state distribution of (Q_{IVR}, Q) be represented by $p(\cdot, \cdot)$, where $p(q_{IVR}, q)$ is the probability that there are q_{IVR} customers in IVR and q customers in ABS in steady state.

The firm clearly has no incentive to deviate from the Full Control solution. Thus, to ensure equilibrium with influential cheap talk we need to ensure that the customers are incentive compatible with respect to the following strategy: the firm provides two distinct signals to differentiate the region $q < \eta^*(q_{IVR})$ from the region where $q \geq \eta^*(q_{IVR})$; and the customers receiving these signals join in the former and balk in the latter. To ensure that the above strategies satisfy Condition 1 of Definition 5.1, we need to ensure that the following two conditions hold:

$$\sum_{q_{IVR}=0}^{\infty} \sum_{q=N}^{\eta(q_{IVR})-1} \left[R - c \frac{q+1-N}{N\mu} \right] p(q_{IVR}, q) \geq 0, \quad (11)$$

$$\sum_{q_{IVR}=0}^{\infty} \sum_{q=\eta(q_{IVR})}^{\infty} \left[R - c \frac{q+1-N}{N\mu} \right] p(q_{IVR}, q) \leq 0, \quad (12)$$

These conditions require that under the firm's full-control solution, if the firm signals "Low Congestion" when $q < \eta^*(q_{IVR})$ and "High congestion" when $q \geq \eta^*(q_{IVR})$, the customer has no incentive to deviate, both when getting the signal that prescribes "join," i.e., $y(m) = 1$ (condition (11)), and when getting the signal that prescribes "balk," i.e., $y(m) = 0$ (Condition (12)). The above conditions can be restated as $R \geq \frac{c}{N\mu} \mathbb{E}[(Q+1-N)^+ | Q < \eta^*(Q_{IVR})]$ and $R < \frac{c}{N\mu} \mathbb{E}[(Q+1-N)^+ | Q > \eta^*(Q_{IVR})]$. Recall that x^+ denotes $\max\{0, x\}$. We define the following two thresholds in terms of the threshold function $\eta^*(\cdot)$ and the

steady state probability function p (but not in terms of the customers' characteristics quantified by R_{IVR} and c_{IVR}):

$$\underline{\eta}^c = \frac{\sum_{q_{IVR}=0}^{\infty} \sum_{q=N}^{\eta(q_{IVR})-1} (q+1-N)p(q_{IVR}, q)}{\sum_{q_{IVR}=0}^{\infty} \sum_{q=N}^{\eta(q_{IVR})-1} p(q_{IVR}, q)}, \quad \bar{\eta}^c = \frac{\sum_{q_{IVR}=0}^{\infty} \sum_{q=\eta(q_{IVR})-1}^{\infty} (q+1-N)p(q_{IVR}, q)}{\sum_{q_{IVR}=0}^{\infty} \sum_{q=\eta(q_{IVR})-1}^{\infty} p(q_{IVR}, q)}. \quad (13)$$

The condition can thus be represented as follows:

$$\underline{\eta}^c \leq \frac{RN\mu}{c} \leq \bar{\eta}^c \quad (14)$$

Notice that the two thresholds $\underline{\eta}^c$ and $\bar{\eta}^c$ are the expected number of customers in the ABS given that the firm wants the customers to join and balk the system, (and given that the customer has to wait in line for the ABS) respectively. Informally speaking, the $\underline{\eta}^c$ is the ‘‘average’’ of the area under the translated threshold function $\eta^*(\cdot) + 1 - N$ weighted with respect to steady state probability of the number of people in the ABS. Similarly, the $\bar{\eta}^c$ defined analogously with respect to the area above $\eta^*(\cdot) + 1 - N$. The key idea behind these definitions is the fact when customers make their join versus balk decisions, they use their belief regarding the number of customers in the ABS and disregard any information or belief regarding the IVR. Thus, we can transform the function η^* into the two relevant thresholds $\underline{\eta}^c$ and $\bar{\eta}^c$.

Note that an equilibrium has to satisfy not only the above conditions, but also Condition 2 of Definition 5.1, which will also demonstrate the dependency on the utility obtained from the IVR. The next Theorem formally presents the conditions for the existence of an equilibrium with influential cheap talk.

Theorem 5.2 *The delayed announcement game has a pure strategy MPBNE with influential cheap talk if and only if*

$$\max \left\{ \frac{\underline{\eta}^c + K}{(1 + \delta(\eta))}, \underline{\eta}^c \right\} \leq \frac{RN\mu}{c} \leq \bar{\eta}^c,$$

where

$$\delta(\eta) = \frac{\sum_{q_{IVR}=0}^{\infty} \sum_{q=0}^{N-1} p(q_{IVR}, q)}{\sum_{q_{IVR}=0}^{\infty} \sum_{q=N}^{\eta(q_{IVR})-1} p(q_{IVR}, q)} \text{ and } K = \frac{(c/\mu_{IVR} - R_{IVR})N\mu}{c \sum_{q_{IVR}=0}^{\infty} \sum_{q=N}^{\eta(q_{IVR})-1} p(q_{IVR}, q)}. \quad (15)$$

Further, $\bar{\eta}^c \leq \sup_q \eta^*(q) - N + 1$.

Proof: The firm clearly has no incentive to deviate from the Full Control solution. Thus, to ensure equilibrium with influential cheap talk we need to ensure that conditions 1 and 2 of the Definition 5.1 hold. Using the arguments presented above we have that condition 1 is satisfied, if and only if, the inequalities in (14) holds. For Condition 2 which requires that the overall expected utility of customers is non-negative, we need to satisfy the following condition:

$$R \sum_{q_{IVR}=0}^{\infty} \sum_{q=0}^{\eta(q_{IVR})-1} p(q_{IVR}, q) - \frac{c}{N\mu} \sum_{q_{IVR}=0}^{\infty} \sum_{q=0}^{\eta(q_{IVR})-1} (q+1-N)^+ p(q_{IVR}, q) + R_{IVR} - \frac{c}{\mu_{IVR}} \geq 0, \quad (16)$$

Using the definition of threshold $\underline{\eta}^c$, $\delta(\eta)$, and K , we can re-express this condition as follows:

$$\left(\frac{c}{\mu_{IVR}} - R_{IVR} \right) + \frac{c}{N\mu} \underline{\eta}^c \sum_{q_{IVR}=0}^{\infty} \sum_{q=N}^{\eta(q_{IVR})-1} p(q_{IVR}, q) \leq R \sum_{q_{IVR}=0}^{\infty} \sum_{q=0}^{\eta(q_{IVR})-1} p(q_{IVR}, q).$$

We can then restate the above condition as follows:

$$\frac{\underline{\eta}^c + K}{(1 + \delta(\eta))} \leq \frac{RN\mu}{c}, \quad (17)$$

where K and δ are as defined in (15). Note that the system dynamics under full control dictate that $Q_{ABS}(t) \leq \sup_q \eta^*(q)$. Thus, we have $\bar{\eta}^c \leq \sup_q \eta^*(q)$. Combining the three inequalities in (14) and (17), we have the desired result. This completes the proof. \blacksquare

Based on the definitions, $\delta(\eta)$ is the ratio of the probability of a customer getting served by the ABS and not experiencing any wait and the probability that an arriving customer gets served by ABS and experiences non-zero wait. Hence δ is strictly positive. The constant K is independent of R and its sign depends on the net utility a customer obtains from the IVR, which equals $R_{IVR} - c/\mu_{IVR}$.

The theorem implies that customers join the system as long as the reward to cost ratio in the ABS is not too high (which entails extreme patience) or too low (which entails extreme impatience). In the setting where the customers receive a high value from the system, they are too patient and are interested in joining a very congested system. Thus, since the customer knows that the firm would like him to balk in a less congested system than the one he would like to join, the signal prescribing “balk” is non-credible and the firm cannot deter the customer from joining. Similarly, when the reward to cost ratio is low, the signal prescribing “join” is non-credible since the firm is interested in luring customers to a more congested system than the one they would like to join.

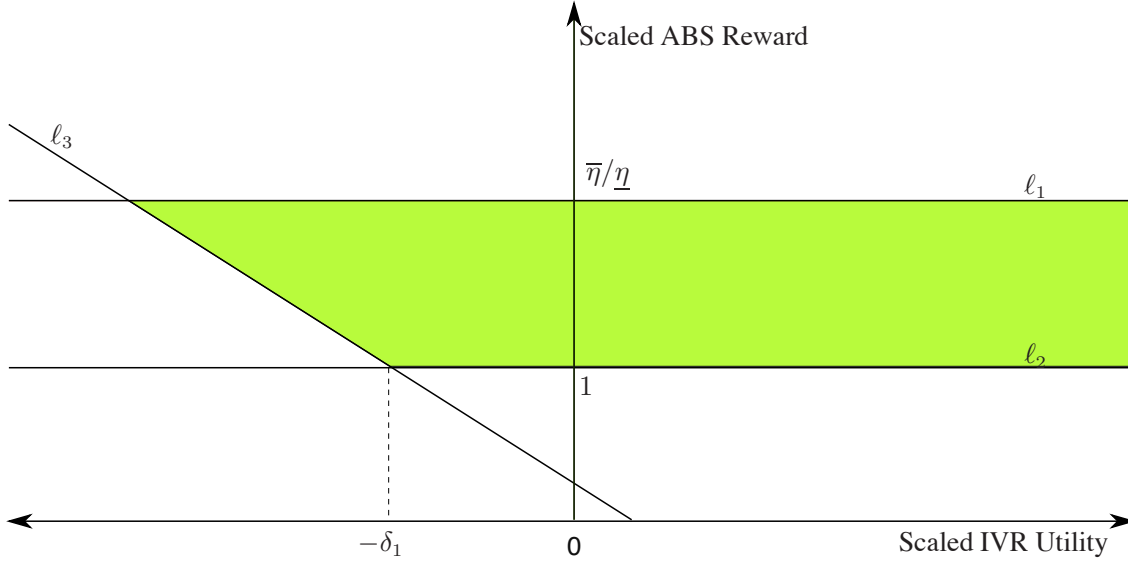


Figure 1 Region of existence of equilibrium with influential cheap talk as a function Scaled ABS Reward ($RN\mu/c\eta$) and Scaled IVR Utility ($(R_{IVR} - c/\mu_{IVR})N\mu/c\eta$).

To understand the impact of the IVR on the threshold we fix the system parameters and vary only the rewards R_{IVR} and R .

In Figure 1, the line ℓ_1 denotes the condition that $RN\mu/c \leq \bar{\eta}^c$. Similarly, the line ℓ_2 denotes the condition that $RN\mu/c \geq \underline{\eta}^c$. The third line (with negative slope) ℓ_3 denotes the condition that $RN\mu/c \geq \frac{\eta^c + K}{1 + \delta}$. Thus the shaded area denotes the set of parameters for the value obtained from ABS, R and the value obtained from the IVR, R_{IVR} for which an MBPNE exists with influential cheap talk. The constant δ_1 denotes $\sum_{q_{IVR}=0}^{\infty} \sum_{q=0}^{N-1} p(q_{IVR}, q)$ which is the probability of a customer getting served without experiencing any delay at the ABS. It is interesting to note that even if the value obtained from the IVR is negative, the firm can still create language that is credible and hence influential. However, as the net-utility obtained from the IVR diminishes, the ability to sustain an equilibrium diminishes, since the value obtained from the system overall decreases.

Note that as long as the net utility from the IVR is not too negative, the value from the IVR has no impact on the ability to credibly communicate delay information. This is formalized in the next corollary.

Corollary 5.3 *Assume that the utility from the IVR is greater than a threshold, i.e., $R_{IVR} - c/\mu_{IVR} \geq -\underline{\eta}^c N\mu \frac{\delta_1}{c}$. Then the delayed announcement game has a pure strategy MPBNE if and only if*

$$\underline{\eta}^c \leq \frac{RN\mu}{c} \leq \bar{\eta}^c. \quad (18)$$

The main implication of the above corollary is that the value customers obtain from the IVR, R_{IVR} , is immaterial to the existence of equilibrium as long as the value obtained from the IVR is not too negative. Note that this stems from the fact that the condition (18) for the existence of equilibrium does not depend on R_{IVR} . Further, the average time spent in waiting for the completion of the IVR processing, i.e. μ_{IVR} , does impact whether an equilibrium exists or not as the thresholds $\underline{\eta}^c$ and $\bar{\eta}^c$ depend on the μ_{IVR} via η^* .

Note also that there exists a threshold such that if the utility from IVR is below this threshold then there is no equilibrium that has influential cheap talk. This nonexistence holds irrespective of the reward that the customers obtain from the ABS. The intuition behind this negative result is as follows: either the customer is not interested in the system at all, due to the large disutility at the IVR, or the firm cannot voluntarily make him leave the system if he is interested (due to a large utility from the ABS).

The next result shows the uniqueness (in terms of the outcomes for the players and system dynamics) of the equilibrium for the above cheap talk game.

Proposition 5.1 *If a pure strategy MPBNE exists, then the MPBNE is unique in terms of the firm's profit, the utility obtained by the customers as well as the dynamics of the system.*

To summarize the two main results of this section, we show that if the condition stated in Theorem 5.2 is violated, then there is no pure strategy MPBNE with influential cheap talk for the delayed announcement game. Thus the condition in this theorem is necessary and sufficient for the existence of pure strategy MPBNE with influential cheap talk. Further, if a pure strategy MPBNE exists, it is unique (in terms of the outcomes) and the firm attains its first best under this equilibrium. With conditions for the existence of equilibrium with influential cheap talk in both games at our disposal, we can now explore the impact of delaying the delay announcement on the ability of the firm to credibly communicate the state-of-the-system.

6. Contrasting the equilibrium strategies in the delayed information cheap talk with the base model

In this section, we shall contrast the equilibria emerging in the delayed cheap talk game and the base game. We would initially study if delaying the announcement of information enhances or detracts from

the possibility of credibly communicating unverifiable information. That is, we would explore if the firm gains or loses the ability to influence customer behavior when the information provision is delayed. From Theorem 5.2, we have $\bar{\eta}^c \leq \sup_q \eta(q) - N + 1$. Recall that $\bar{\eta}^B = \hat{q} - N + 1$, we derive the following bounds on $\sup_q \eta(q)$ and $\bar{\eta}^B$ that would be useful for this study.

Proposition 6.1 *We have the following*

- (a) *For the base model, $\bar{\eta}^B \leq vN\mu/h$.*
- (b) *For the delayed cheap talk model, $\bar{\eta}^c < \sup_q \eta(q) - N + 1 \leq vN\mu/h$.*

In both the base model as well as the delayed cheap talk model, we observed that if the customers are extremely patient then the firm cannot sustain an equilibrium with influential cheap talk. Note that if the customer has full information, they will not join if the number of customers in the system exceeds $RN\mu/c + N - 1$. Thus, $RN\mu/c$ can be viewed as the patience of the customers. In the base model, if $RN\mu/c > \bar{\eta}^B$, the firm cannot support an equilibrium with influential cheap talk, whereas for the delayed cheap talk game the firm cannot support an equilibrium with influential cheap talk if $RN\mu/c > \bar{\eta}^c$. The above results show that both of these thresholds $\bar{\eta}^B$ and $\bar{\eta}^c$ are bounded above by $vN\mu/h$. Thus, if $R/c > v/h$ then there is no equilibrium with influential cheap talk in both of the cheap talk games. This points to the fact if the customers have extreme patience, i.e., $R/c > v/h$, no matter how much foresight is obtained through the IVR by the firm, delaying the announcement will not help the firm in influencing its customers. Thus, for these parameters, the firm does not gain or lose any ability which it had without delaying.

However, if $\bar{\eta}^B < RN\mu/c < \bar{\eta}^c$, then clearly the firm gains by permitting an equilibrium in a region in which it could not communicate information without delaying the information provision. We next demonstrate that the firm can also diminish its capability to communicate credible information to its customers.

Observe that the necessary and sufficient condition for the existence of equilibrium for the delayed cheap talk can also be written as $RN\mu/c \in [\underline{\eta}^c, \bar{\eta}^c]$. Here one can view $RN\mu/c$ as providing the customers' perspective, and $\underline{\eta}^c, \bar{\eta}^c$ as the firm's perspective on the desired congestion level in the system. In studying the impact of delaying the announcement, we shall fix the firm's perspective and vary the customers' perspective. As shown in Proposition 3.1 the necessary and sufficient condition for the existence of equilibrium is

$$RN\mu/c \in [\underline{\eta}^B, \bar{\eta}^B].$$

We introduce the following terminology: fixing the cost parameters for the firm as well as the service rate and arrival rate, let S and S_d denote the set of the customers' thresholds q^* for which the firm can sustain an equilibrium with influential cheap talk without delaying the announcement (in the base model) and with delaying it (in the delayed cheap talk game), respectively. Based on the above discussion, we have that the set $S = [\underline{\eta}^B, \bar{\eta}^B]$ and the set $S_d = [\underline{\eta}^c, \bar{\eta}^c]$. We define the expansion region due to the delayed provision as $S_d \cap S^c$, where S^c denotes the complement of the set S . Similarly, we define the contraction region due to the information provision, as $S \cap S_d^c$ where S_d^c denotes the complement of the set S_d . Lastly, we define the neutral region due to the information provide $S_d \cap S$.

We shall say that delaying the information provision results in a *contraction* if the expansion region is empty. Similarly, we say that delaying the information provision results in an *expansion* if the contraction region is empty. We will say that the information provision results in *mixed contraction-expansion* if neither of these sets is empty. Figure 2 depicts the contraction and expansion region for a case where the information provision results in mixed contraction-expansion.

Based on Proposition 6.1, we have the following immediate corollary.

Corollary 6.1 *The expansion, contraction and neutral regions are subsets of $[0, vN\mu/h]$.*

The main implication of the above corollary is that there is a limit on the extent by which a firm can expand the set on which it provides credible information by delaying the information provision. Further, the set of customers' thresholds on which the firm can improve its credibility is bounded irrespective of how much the firm delays the customer prior to providing the information.

We shall next illustrate the expansion and contraction regions via two numerical examples.

Example 1: For the first example, we assume the arrival rate, $\lambda = 0.5$ customers per unit of time. There is a single agent whose service rates in both the IVR and ABS are assumed to be unity, i.e., $\mu = \mu_{IVR} = 1$. The ABS is staffed with a single agent. We assume that the firm obtains a value of $v = 15$ from each served customer, yet incurs a holding cost of $h = 0.5$ per customer per unit of time. We use a discount rate of $\alpha = 0.05$ for the firm. We evaluate the optimal policy using value iteration over a truncated state-space.

Based on the optimal threshold policy η^* , we compute the thresholds defined in (13) $\bar{\eta}^c = 11.15$, $\underline{\eta}^c = 1.99$. Thus, using Theorem 5.2 we have that an equilibrium with influential cheap talk (i.e., $g(Q_{IVR}, Q) = m_1$ if $Q \leq \eta(Q_{IVR})$ and m_2 otherwise, and the customer uses $y(m_1) = 1$ and $y(m_2) = 0$), exists in the delayed cheap talk model iff $1.99 \leq R/c \leq 11.15$. For the base model, we compute the optimal full-control solution. The optimal threshold for the firm is $\hat{q} = 8$. Under this solution, the average number of customers in the system when an arriving customer is recommended to join the system is $\underline{q} = 0.97$. Thus, for the base model, an equilibrium with influential cheap talk exists iff $0.97 \leq R/c \leq 8$. Note that in this case, the expansion region is $[8, 11.15]$ and the contraction region $[0.97, 1.99]$. That is, if $R/c \in [8, 11.15]$ delaying the delay announcements allows the firm to augment the possibility of credibly communicating delay information to its customers. On the other hand, if $R/c \in [0.97, 1.99]$, delaying the information provision detracts from the possibility of credible communication between the firm and its customers. Thus, in this example, depending on the valuation of the customers, delaying the information provision may augment, detract from or have no impact on the equilibrium language.

The existence of an expansion region above implies that we have a region in which an influential language exists due to the postponement of the delay announcement. The main difference with the base model due to which the language gets augmented is that now the customer cannot detect the state of the system when the firm suggests that he balks. By doing that, the vagueness in the announcement increases, enabling credible communication where it was not possible in the absence of such postponement. Thus, the firm not only gains due to extra information (the state of the IVR) but also gains due to the vagueness it can create.

Example 2: While intuitively, one may expect delaying the information to always augment some of the language, we next show that this is not always the case. We use the same parameters as in Example 1, with the following modification: $\lambda = 0.97$. Based on the optimal threshold policy η^* , we compute the thresholds defined in (13) $\bar{\eta}^c = 6.76$, $\underline{\eta}^c = 3.58$. Thus, using Theorem 5.2 we have that an equilibrium with influential cheap talk exists in the delayed cheap talk model iff $3.58 \leq R/c \leq 6.76$. For the base model, we compute the optimal full-control solution. The optimal threshold for the firm is $\hat{q} = 7$. Under this solution, the average number of customers in the system when an arriving customer is recommended to join the system is $\underline{q} = 2.55$. Thus, for the base model, an equilibrium with influential cheap talk exists

iff $2.50 \leq R/c \leq 7$. In contrast to the previous example, note that in this case there is no expansion region. The contraction region consists of two intervals $[2.55, 3.58]$ and $[6.76, 7]$. That is, if $R/c \in [2.55, 3.58]$ or $R/c \in [6.76, 7]$ delaying the information provision detracts from the ability of credible communication between the firm and its customers. Thus, in this example, depending on the valuation of the customers, delaying the information provision may either detract from or have no impact on the equilibrium language. Hence, the firm cannot gain credibility by delaying the delay announcement. Note that in this case, the utilization in the system is higher compared to the one in Example 1. Due to this high utilization, the externalities a customer imposes on other customers play a more crucial role in the firm's decision whether to accept him or not. However, in equilibrium, the customers know this fact, and "resent" the greedy nature of the firm, and thus the firm loses its credibility in some regions.

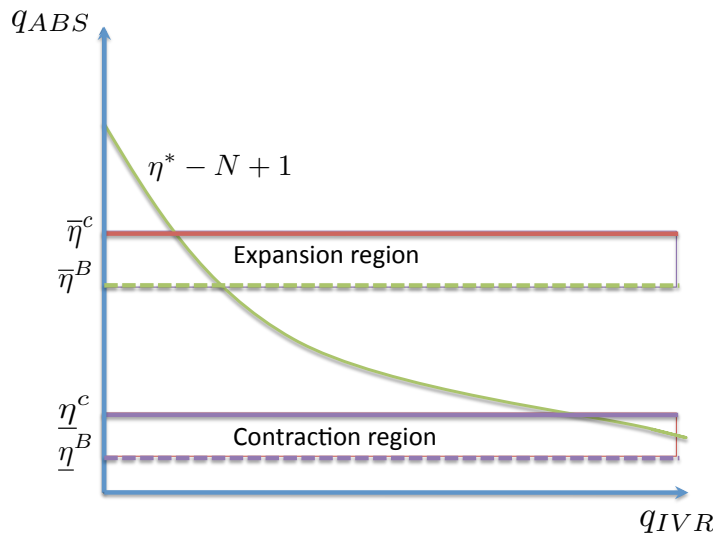


Figure 2 Expansion and contraction regions. The above figure depicts a setting where there are both an expansion region and a contraction region.

Based on the above examples, we observe that delaying the announcement may enhance the possibility for information transmission, but also may hamper its possibility. The former occurs as the firm can create more vagueness of the mapping between the state of the system and the announcements. The contraction, which is more surprising, occurs due to the firm's attempt to exploit the additional information to generate higher

profits, which might increase the misalignment between the firm and the customer. The above discussion focused on the impact of delaying the announcement on the emerging equilibrium language. Next, we shall study whether delaying the information provision translates into improved outcomes for the different parties.

6.1. The value of delay

In the above discussion, we studied the ability of the firm to gain credibility with regard to the delay announcements. Moreover, it is important to understand if this creation of credibility translates into value creation for the firm and its customers. While the discussion on the impact of delaying the delay announcement on the ability of the firm to improve its credibility required making no assumptions on the value or cost of waiting in the IVR, in order to understand the value created due to this postponement, we would like to focus on the common part of service which is the ABS. To do so, we will make the assumption that neither the firm nor the customer creates any value from the IVR. This assumption is relaxed at the end of this section. Section 7.2 also discusses the case of a system with essential IVR. In order to accomplish this we assume that customers obtain zero net-utility from the IVR, that is $R^{IVR} = \frac{c}{\mu_{IVR}}$, and the firm obtains nothing from the IVR directly, that is, $v^{IVR} = \frac{h}{\mu_{IVR}}$. Under this setup, we have the following result.

Proposition 6.2

- (a) *For any $RN\mu/c$ that belongs to the expansion region, both the customers and the firm would be better off in the delayed cheap talk game.*
- (b) *For any $RN\mu/c$ that belongs to the contraction region, both the customers and the firm would be better off in the base model compared to the delayed cheap talk game.*
- (c) *For any $RN\mu/c$ that belongs to the neutral region, both the customers and the firm would be better off under the equilibrium with influential cheap talk for the delayed cheap talk game compared to any equilibrium for the base model.*

The proposition shows that when the customers' threshold belongs to the expansion or neutral regions, delaying the information provision allows both the firm and the customers to improve their profits and utilities, respectively. On the other hand, if the customers' threshold belongs to the contraction region,

their respective profits and utilities would diminish. Note that the above proposition uses the fact that in the delayed cheap talk game the customers and the firm are better off in an equilibrium with influential cheap talk. Since the firm improves its profits going from the base model to the delayed cheap talk, one may expect these profits to come at the expense of the customers, as the firm lures customers in states they would otherwise not join, or turns away the customers in states they would have joined if given full information. However, we show above that the customers, together with the firm, enjoy the augmentation of the equilibrium language and suffer from its contraction, see Theorem 5.2.

It is important to note that our results are consistent with recent results in the literature exploring the value of postponement in supply chain models. For example, Anand and Girotra (2007) demonstrate that purely strategic considerations play a pivotal role in determining the value of delayed differentiation in a supply chain configuration. They show that in the face of either entry threats or competition, these strategic effects can significantly diminish the value of delayed differentiation, and that these effects may dominate the traditional risk-pooling benefits associated with delayed differentiation.

While the net-zero utility of the IVR is assumed above to level the playing field between the two models, it is important to note that the characterization of the equilibrium using the two thresholds, as well as the expansion/contraction regions hold regardless of the utility (disutility) customers obtain during their stay at the IVR, as long as the utility is not too low.

Note however, if the customer's overall utility from the IVR is negative, one may encounter situations in which delaying the delay announcement improves the overall profit of the firm (as long as the value the firm obtains is not negative), yet reduces the utility for the customer.

7. Extensions

In this section, we extend our findings in two directions. First, we allow the customers to abandon the system, i.e., the customers, once they join, are not required to get served. Second, we study a related model where the IVR is essential to the system and the customers may be given the delay announcement before or after this essential IVR.

7.1. Abandonments

In many service settings, customers can make a decision not only regarding joining versus balking, but also about leaving the system after joining without receiving the service. So far, we have focused on the first two decisions, while disallowing customer abandonment. In this section we first show that in the base model, where the information is provided immediately upon arrival, if the customers are allowed to update their beliefs about the system and renege the queue, the equilibria characterized remain unchanged. We then provide conditions under which a similar property is exhibited by the delayed cheap talk model.

Proposition 7.1 *In the equilibria identified for the base model in Proposition 3.1 a rational customer will not abandon even if allowed to, and the firm will not deviate from its signaling rule. In this sense, these equilibria are abandonment-proof.*

The proposition states that a rational customer who updates his belief on the state of the system *after* joining the system, would not abandon. While the customer might realize the fact that he was lured to join in a state he otherwise would not join, he is in a better position compared to the one in which he decided to join. This is somewhat equivalent to treating the elapsed time as “sunk cost,” and explains the reason why the customers do not have any profitable deviation. On the other hand, one can easily see that given that the customers who join the system are not interested in leaving the system without service, the firm’s full control solution remains unchanged. Thus, the firm will not have any profitable deviation either.

Hence, in this setting abandonments will not arise endogenously. Other more complex settings, such as the one in which the valuation varies over time (see Haviv and Ritov 2001) or one in which the customers feel that they have been left out of the system without being informed (see Mandelbaum and Shimkin 2000), can lead to rational abandonments.

Next, we turn our attention to the delayed cheap talk game. Let Q denote the number of customers in the queue in ABS at the departure instance from the IVR, conditioned on the fact that he joins the ABS and waits for his service. It is clear that in the delayed model, using the above logic that a customer will not abandon during his stay at the IVR. It is worth noting that this abandonment-proofness for IVR stems from the fact that the delay in the IVR is independent of the congestion in the system. Thus, the customer

is not able to obtain any additional information about the system as he gets “served/delayed” at the IVR. The question of whether one will abandon while waiting for the ABS can be answered along the same lines as the base system. The following result provides sufficient condition for the abandonment-proofness of the equilibria defined for the delayed cheap talk game.

Proposition 7.2 *If Q has an increasing hazard rate under the equilibria defined by Proposition 5.2 then a rational customer will not abandon even if allowed to in the IVR as well as in the ABS and the firm will not deviate from its signaling rule. In this sense, these equilibria are abandonment-proof.*

7.2. The case of an essential IVR

In this section, we consider a system where the IVR provides an essential part of the service. Specifically, we study the following two setups that include the IVR: a) a system where the firm provides information immediately on arrival (before the IVR), and b) a system where the firm provides delayed delay announcement (after the IVR). Given that the IVR is essential, we would assume that both the customer and the firm derive positive utility from the experience at the IVR, i.e., $v_{IVR} > h_{IVR}/\mu_{IVR}$ and $R_{IVR} > c_{IVR}/\mu_{IVR}$.

Using an arguments similar to Proposition 4.1, if the firm has full control in the system with immediate delay announcement (i.e. announcing the delay before the IVR) the optimal policy has the following property.

Proposition 7.3 *For the full control problem in the system with immediate delay announcements, there exists a threshold function $\tilde{\eta}^*(\cdot)$, such that it is optimal for the firm to accept a customer that completes service at the IVR, if and only if $Q < \tilde{\eta}^*(Q_{IVR}(t))$ and “turn him away” otherwise.*

Proof: To show that $V_{IA}^\kappa(q_{IVR}, q)$ is concave in q for a given q_{IVR} we follow the same lines as the in proof of Proposition 4.1. To show that this holds for $V(q_{IVR}, q)$, we have to use a slightly different argument. In particular, it is easy to see that

$$V_{IA}^\kappa(q_{IVR}, q) < V_{IA}(q_{IVR}, q) < V_{IA}^\kappa(q_{IVR}, q) + (v_{IVR} + v)\mathbb{P}(Q_{M/M/\kappa/\kappa} = \kappa).$$

Then, taking limits as $\kappa \rightarrow \infty$ we get the desired result. ■

Further, under full control we can show that for any initial state, the system with immediate announcement performs worse than the system with delayed announcement. To formally state the result, let $V_{IA}(q_{IVR}, q)$ and $V_{DA}(q^{IVR}, q)$ be the optimal discounted infinite horizon profit in the immediate announcement (IA) system and the delayed announcement (DA) system starting with the state (q_{IVR}, q) , respectively. Then we have the following result.

Proposition 7.4 *We have*

$$V_{IA}(q_{IVR}, q) \leq V_{DA}(q^{IVR}, q).$$

We can define the notion of pure strategy MPBNE for the cheap talk games for both the immediate announcement as we have done for the delayed announcement system. Recall that an equilibrium with influential cheap talk exists in the delayed announcement system iff the conditions specified in Theorem 5.2 hold. Similarly, one can observe that the influential influential cheap talk exists in the immediate announcement system iff the following conditions hold:

$$\sum_{q_{IVR}=0}^{\infty} \sum_{q=N}^{\eta(q_{IVR})-1} \mathcal{U}(q_{IVR}, q)p(q_{IVR}, q) \geq 0, \quad \sum_{q_{IVR}=0}^{\infty} \sum_{q=\eta(q)}^{\infty} \mathcal{U}(q_{IVR}, q)p(q_{IVR}, q) \leq 0, \quad (19)$$

where $\mathcal{U}(q_{IVR}, q)$ is the expected utility obtained by a customer from the system (including IVR) joining the immediate announcement system where the firm has full control and the state of the system is (q_{IVR}, q) . Using Proposition 7.4 we obtain the following result.

Corollary 7.1 *If there exists a pure strategy MPBNE for both the immediate announcement cheap talk game as well the delayed announcement cheap talk game, then the firm would earn higher profits in the latter.*

The implication of this theorem is that if credibility can be obtained via delaying the delay announcement or sustained in both systems regardless of the timing of the announcement, the firm prefers the setting in which the information provision is delayed. However, it is important to note that, just as in the comparison discussed in Section 6, while the firm might expand the set of equilibria by delaying the information provision, the firm might also diminish the set.

8. Conclusion

Many service providers as well as make-to-order manufacturers use delay announcements to inform customers on the level of congestion in the system, usually not providing the information immediately, but rather after a short period of time (spent either waiting or occupied by the system). The focus of this paper is on the impact of this postponement on the ability of the firm to communicate non-verifiable congestion information to its customers as well as on the profits and utilities for the firm and the customers, respectively.

It is clear that if the firm has full control, delaying the announcement allows the firm to improve the profit it obtains from the agent-based-service. This is due to the fact that the firm has better knowledge of the externalities customers impose on the system when the firm makes its admission decision for the ABS. However, in practice, it is difficult and also very expensive to ask a customer to leave once he is admitted to the system. In this paper, we show that under certain settings, this optimal admission control policy can be achieved by providing delay announcement. In fact, this delay can actually help the firm create credibility and augment the equilibrium language (using the additional level of vagueness). However, this delay can also detract from the equilibrium language (given that the firm is more sophisticated in its strategies, it can hurt its credibility). Further, we show that whenever credibility is created it improves not only the profit for the firm but also the customers' overall utility, provided the net utility from the IVR is not too negative.

Future research. It is worth exploring empirically the validity of the above findings. While many firms provide information to their customers, to the best of our knowledge, there is no empirical study that shows how and if the customers react to these announcements. The idea of delaying the delay announcements can be studied also in a retail setting. For example, a customer may be notified about the availability of an item during the pre-season period. The firm may use this information to obtain advanced-sales information and may inform the customer about the likelihood of stocking out.

References

- Allon, G., A. Bassamboo, I. Gurvich. 2007. "We will be right with you": Managing customers with vague promises and cheap talk. *Working paper, Kellogg School of Management, Northwestern University*.
- Armony, M., C. Maglaras. 2004a. Contact centers with a call-back option and real-time delay information. *Oper. Res.* **52**(4) 527–545.

- Armony, M., C. Maglaras. 2004b. On customer contact centers with a call-back option: customer decisions, routing rules and system design. *Oper. Res.* **52**(2) 271–292.
- Armony, M., N. Shimkin, W. Whitt. 2009. The impact of delay announcements in many-server queues with abandonment. *Operations Research* **57**(1) 66–81.
- Crawford, V. P., J. Sobel. 1982. Strategic information transmission. *Econometrica* **50** 1431–1451.
- Feigin, P. 2006. Analysis of customer patience in a bank call center. Tech. rep., Working paper in preparation, The Technion.
- Ghoneim, H.A., S. Stidham. 1985. Control of arrivals to two queues in series. *European Journal of Operational Research* **21** 399–409.
- Guo, P., P. Zipkin. 2007. Analysis and comparison of queues with different levels of delay information. *Management Science* **53**(6) 962–970.
- Hassin, R. 1986. Consumer information in markets with random product quality: The case of queues and balking. *Econometrica* **54** 1185–1195.
- Haviv, M., Y. Ritov. 2001. Homogeneous customers renege from invisible queues at random times under deteriorating waiting conditions. *Queueing Systems* **38**(4) 495–508.
- Hui, M.K., D.K. Tse. 1996. What to tell consumers in waits of different lengths: an integrative model of service evaluation. *The Journal of Marketing* **60**(2) 81–90.
- Ibrahim, Rouba, Ward Whitt. 2008. Real-Time Delay Estimation Based on Delay History. *Forthcoming in MSOM*.
- Jouini, O., Y. Dallery, O.Z. Aksin. 2007. Queueing models for multiclass call centers with real-time anticipated delays. *International Journal of Production Economics*, To Appear.
- Koole, G. 1998. Structural results for the control of queueing systems using event-based dynamic programming. *Queueing Systems* **30**(3) 323–339.
- Ku, C.Y., S. Jordan. 2002. Access control of parallel multiserver loss queues. *Performance Evaluation* **50**(4) 219–231.
- Ku, C.Y., S. Jordan. 2003. Near optimal admission control for multiserver loss queues in series. *European Journal of Operational Research* **144**(1) 166–178.
- Kumar, P., M.U. Kalwani, M. Dada. 1997. The impact of waiting time guarantees on customers' waiting experiences. *Marketing Science* **16**(4) 295–314.
- Kumar, P., P. Krishnamurthy. 2008. The Impact of Service-Time Uncertainty and Anticipated Congestion on Customers' Waiting-Time Decisions. *Journal of Service Research* **10**(3) 282.
- Mandelbaum, A., N. Shimkin. 2000. A model for rational abandonments from invisible queues. *Queueing Systems* **36**(1) 141–173.
- Ross, S.M., J.G. Shanthikumar, Z. Zhu. 2005. On increasing-failure-rate random variables. *Journal of Applied Probability* **42**(3) 797.

Spence, A. M. 1973. Job market signaling. *Quarterly Journal of Economics* **87** 355–374.

Whitt, W. 1999. Improving service by informing customers about anticipated delays. *Management Science* **45** 192–207.

Appendix A: Proofs

Proof of Proposition 5.1 For the above delayed cheap talk game, if an informative equilibria exists, it must be the case that the firm obtains its first best. This can be argued as follows: suppose there exists an informative equilibrium where the firm does not obtain its first best. It must be the case that there are at least two signals which are used by the firm and the customer joins when she receives one signal and balks when she receives the other signal. Given this strategy for the customer, the firm would have a profitable deviation if it does not achieve its first best. Thus, we have the result. ■

Proof of Proposition 6.1 For part (a) note that the contribution of a customer to the profit of the firm is $v - hW$, where W is the waiting time in the system. Clearly, if the contribution of the customer is negative, the firm will not admit him. Note, however, that due to the need to account for the dynamics of the model, a customer with positive contribution is not necessarily guaranteed admittance. Thus, if the number of customers in the system exceeds $v\mu/h$, the expected waiting time would be greater than v/h , thus his contribution will be negative. This completes the proof of part (a). Proof of part (b) is analogous to the above proof, and uses the observation that if a customer provides only negative contribution based on the number of customers in the ABS, the firm can disregard the number of customers waiting in the IVR. ■

Proof of Proposition 6.2

(a) For an informative equilibrium to exist it must be the case that $R/c < v/h$. Further, the firm achieves its first best profit. Also, the only pure strategy non-informative equilibrium for the system is the one where no customer balks the system. Let $\mathbb{E}[\widetilde{W}]$ be the waiting time in this system. Also, let p_0 denote the fraction of customers who are joining in an informative equilibrium for the delayed cheap talk game. Let $\mathbb{E}[W]$ denote the expected waiting time in ABS for the system in an informative equilibrium for the delayed cheap talk game. Using the fact that the firm achieves its first best under an informative equilibria, we have:

$$\lambda(v - h\mathbb{E}[\widetilde{W}]) \leq \lambda p_0(v - h\mathbb{E}[W]).$$

The above implies

$$\frac{v}{h} \leq \frac{\mathbb{E}[\widetilde{W}] - p_0\mathbb{E}[W]}{1 - p_0}.$$

Appealing to the fact that $R/c < v/h$, we have that

$$\lambda(R - c\mathbb{E}[\widetilde{W}]) \leq \lambda p_0(R - c\mathbb{E}[W]).$$

Thus, the overall expected utility of the customers would improve in an informative equilibrium when compared to a non-informative one. Noting that the babbling equilibrium in the base model and the delayed cheap model are identical, completes the proof.

(b) Combining Propositions 4.5 and 5.1 from Allon et al. (2007), we obtain that in the base model the customers and the firm prefer an informative equilibrium over a babbling one. Further, note that the non-informative equilibrium in the base model and the delayed cheap model are identical. This completes the proof.

(c) The proof follows in an analogous manner to part (a) with the exception that there will be a probability of joining for both the base model and delay announcement model. Further, since $R/c < v/h$ for the neutral region, we obtain the result. ■

Proof of Proposition 7.1 Note that under the proposed equilibrium the system operates as an $M/M/N/k$ queueing system. We next show that the customers who wait in this queueing system experience waiting time with increasing hazard rate.

In an $M/M/N/k$ system using the steady state analysis and the PASTA property, the number of customers waiting in the queue when a customer arrives and joins the system (and waits) has a geometric distribution conditioned on it being less than k . Thus, we have that the waiting time in the system for a customer has the same distribution as $\sum_{j=1}^{Q+1} X_j$, where X_j are exponential with mean $1/(N\mu)$ and Q has the same distribution as the number in system of $M/M/N/k$ conditioned on the event that the number of queue is less than N and the arriving customer waits. We also know that the $Q + 1$ has an increasing hazard rate. Using Theorem 7.1 in Ross et al. (2005), we have that the waiting time will also have an increasing hazard rate.

Using the arguments similar to the proof of Proposition 2(i) in Mandelbaum and Shimkin (2000), we obtain that the customers will never abandon the system if they join the system. This shows that the equilibria described earlier are abandonment-proof from the customer's perspective. Given that the customers do not abandon and the firm had no profitable deviation from its signaling rule, it follows that the firm will not have any profitable deviation when the customers are allowed to deviate. This completes the proof. ■

Proof of Proposition 7.2 The proof follows by appealing to Theorem 7.1 in Ross et al. (2005) and Proposition 2(i) in Mandelbaum and Shimkin (2000), and is along the same lines as Proposition 7.1. ■

Proof of the Proposition 7.4 Consider the two full control problem on the same probability space. All the service times and arrival times for the two systems are identical. Consider the optimal policy of the full control problem under immediate announcement. It is easy to see that if we employ a policy for the delayed announcement that rejects exactly those customers rejected in the immediate announcement system then the ABS part of the system behaves in an identical manner for the two systems. Moreover, since we have $v_{IVR} > h/\mu_{IVR}$, the delayed announcement system generates more profit. Thus, we have the desired result. ■