

# 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

December 1, 2008

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 communicate 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 equilibrium language. However, in other settings this delay can also detract the equilibrium language. Further, we show that whenever credibility is created it improves not only the profit for the firm, but also the customers' overall utility.

---

## 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 provides them with the congestion level the system is currently experiencing. These announcements can improve the customer service experience and the system performance. Most service organizations provide such information only after some waiting 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 the firms to credibly communicate delay information to its 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 some are as vague as the statement “the system is experiencing long waits” to signal to the customers the congestion the system is facing. In these cases, even after the firm categorizes the needs of the customer, it delays providing the waiting time information to the customers. These delays can be inserted using many mechanisms such as providing current promotions

information or rebates available (for example, 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) or providing the information about other channels which the customers can use (for example, used by banks such as Washington Mutual, and by many airlines to divert customers to their websites by informing them of these options).

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 signal, or by diverting customers immediately 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 him of the 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 postponing 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 costless<sup>1</sup>, 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

<sup>1</sup> 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 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 consider a setting where the service process is comprised of two components. These are modeled as a tandem queue. The first part of the process has ample capacity so no customer is in the queue, in fact every arriving customer enters the service immediately on arrival. Thus this component models IVR for call center setting. For the second component there is finite capacity and thus the customer may experience delays if the system is congested. We study the impact of delaying the delay announcement by comparing two settings: one where IVR is present and the customer is given the delay announcement after he is processed through IVR, and another where there is no IVR and the customer is provided with information as soon as he joins the system. Specifically, 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 vs. balk.

To analyze the game in both settings (when the announcement is made upfront versus the case where announcement is delayed), one needs to analyze these systems where the firm can dictate the decision of the customers. We refer to these problems as full admission control and full access control for the two settings, respectively. We begin by analyzing the full admission 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 control in the full access control problem. Using these results, we then 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, 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 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 informative equilibrium exists. We show that for an informative equilibrium to exist the ratio between the customer's value of the service to his cost of waiting has to be above a certain threshold level, but below a different one. A similar result applies to the non-delayed cheap talk model where the conditions for existence of an informative equilibrium can be described using two different threshold levels.

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. To do this comparison, we assume that both the customers and firm obtain zero net value from the first stage. In these setting we show there are instances where the firm can create credibility for its messages due to delaying. However, it might also *lose* its credibility due to the delay. We further, show that both the firm and the customers always prefer an equilibrium which is informative (even though it might be created due to the delay and even though the customer is lured to systems in states he would otherwise not join) over the non-informative equilibria. We also discuss cases where the net value of the first stage is nonzero.

*Organization of the paper.* In the next section, we discuss the relevant literature. Section 3 describes and analyzes the base-model where the information is provided upfront and there is no delay. Section 4 describes the tandem queue model and analyzes the full access control problem. Section 5 describes and analyzes as the delayed cheap talk game. Section 6 contrasts the equilibrium strategies as well the outcomes. This section also provides numerical study. Section 7 provides the conclusion to the paper.

## 2. Literature Review

*Delay announcement models.* There is a growing interest in models studying the impact of delay announcement, when the queue is invisible, on the system performance. 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. Armony and Maglaras (2004) extends the above model to allow the service manager to provide the customers an estimate of the delay, based on the state of the system upon their arrival. Armony et al. (2007) 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 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 settings where no information on the queue length is provided, and 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.

The issue of providing customers with delay information arises also in a manufacturing environment where firms quote leadtimes. Duenyas and Hopp (1995) studies the problem of quoting customer lead times along with the optimal control, both with infinite and finite capacity. Ata and Olsen (2007) studies a

related problem for large systems under convex-concave cost structure. Dobson and Pinker (2006) develops a stochastic model of a custom production environment with pricing, where customers have different tolerances for waiting. The authors model intermediate levels of information sharing (with a specific structure) ranging from none to complete state-dependent lead-time information, and compare the performance from the firm's and customers' perspectives. They show that for this specific structure it is not always the case that sharing information improves the profits of the firm.

*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 Jr (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 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 queueing system. The authors prove that the optimal admission control policy is given by a set of thresholds. In these papers, the control policy can only reject arriving customers from outside the system. A customer that was already accepted to the system cannot be blocked by the system manager. (Ku and Jordan (2003) characterizes the least restrictive admission control policy for a tandem loss network such that only external customers are rejected.) This is due to the fact that the cost of blocking an accepted customer is exorbitantly high, compared to blocking a new customer. This is in contrast to our model, where the customers themselves terminate their call or request for service based on *their* assessment of the quality of the service.

*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 queueing 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 leads 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.

All of the above mentioned models of delay announcements have studied settings where information is announced immediately as the customers arrive and communicates some information on the anticipated delays. In this paper, we will consider a setting in which the firm is capable of delaying the delay announcement.

### **3. Base Model: Benchmark**

In this section, we first develop a model where the service provider announces the delay related information immediately to an arriving customer. This case would 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/1 system. Customers arrive to the system according to a Poisson process with rate  $\lambda$ . Service times are exponentially distributed with mean  $1/\mu$ . We assume that  $\lambda < \mu$ . 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 sojourn time in the system. Throughout the paper, we shall assume that  $R > \frac{c}{\mu}$ , this assumption ensures that in the absence of delays, the service is beneficial to the customer, on average. Clearly, if  $R < \frac{c}{\mu}$ , no customer will join regardless of the system announcements. 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, denoted by  $I$ , in order to maximize its expected utility. Therefore the customer will join, if and only if  $R \geq c\mathbb{E}(w|I)$ , where  $I$  is the information provided to this customer. The system manager obtains revenue of  $v$  per customer served, and incurs a holding cost  $h$  per unit of time per customer. The firm's profits are then given by the following expression:

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

where  $D(t)$  is the departure process from the system.

We next define the notion of Markov Perfect Bayesian Nash Equilibrium (MPBNE) for the cheap talk game played between the customers and the firm. In practice, one observes different types of messages that are conveyed to the customers. These messages could provide tangible information such as expected wait in the queue or the position of the customer in the queue. Also, in some cases, the message may only have intangible information such as the congestion level or volume of calls being high or low. In this paper, we propose a framework that allows us to study the provision of unverifiable information using a unified approach regardless of the type of the announcement. In this framework we account for the following key features: a) the state-of-world changes dynamically; b) the customer cannot verify the information provided by the firm; and c) the customers would process any information provided to them by the firm to base their

action on it. Now that the structure we lay out covers the suggested framework, we can allow ourselves not to be restricted to any type of announcements and in particular, it allows us to treat announcements of the variety mentioned above.

Towards the definition of the game, 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 queue length 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.

**Definition 3.1 (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. Let  $p_q(y, g)$  be the steady state probability that the number of customers in an  $M/M/1/N$  is  $q^2$ . For each  $m \in \mathcal{M}$ , we have

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

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

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

Following the arguments used in Proposition 3.1 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. This is formalized in the following result whose proof is omitted.

**Proposition 3.1** *Let the pair  $y(m)$  and  $g(q)$  be a pure strategy MPBNE. Then there exists a constant  $\bar{q}$  such that the pair  $(\tilde{g}(\cdot), \tilde{y}(\cdot))$  given by*

$$\tilde{g}(q) = \begin{cases} m_1 & q \leq \bar{q}, \\ m_0 & \text{otherwise.} \end{cases}, \quad \tilde{y}(m) = \begin{cases} 1 & m = m_1, \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

<sup>2</sup> Note that  $p_q(y, g)$  can be thought of as the beliefs of the agents on the state of the systems. These beliefs are consistent with the the strategy of the other players.

forms a MPBNE with the same firm profit and customer utility.

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

$$q^* = \left[ \frac{R\mu}{c} \right], \quad (4)$$

where  $[\cdot]$  is the bracket function; i.e.,  $q^*$  is the largest integer not exceeding  $R\mu/c$ . 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  $M/M/1/k$  queueing system. 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  $k$ , its expected profit is given by

$$\Pi(k) = \mathbb{E} \left[ \int_0^\infty e^{-\alpha t} v dD^k(t) - \int_0^\infty h Q^k(t) dt \right]. \quad (5)$$

The following proposition shows that the profit function is unimodal in  $k$ .

**Proposition 3.2** *The function defined by  $\Pi(k)$  is unimodal in  $k$ , i.e., there exists  $k^* \in \{1, 2, \dots, \infty\}$  such that the function  $\Pi(k)$  is strictly increasing for  $k < k^*$  and strictly decreasing for  $k \geq k^*$ .*

**Proof:** Note that the firm's profit can be expressed as

$$\begin{aligned} \Pi(k) &= \mathbb{E} \left[ \int_0^\infty e^{-\alpha t} v dD^k(t) - \int_0^\infty h Q^k(t) dt \right] \\ &= \frac{v\lambda}{\alpha} (1 - p_k^k) - \frac{h}{\alpha} \sum_{j=1}^k j p_j^k \end{aligned}$$

$$= \sum_{j=0}^{k-1} \left[ \frac{v\lambda}{\alpha} - \frac{h\lambda}{\alpha\mu} j \right] p_j^k,$$

where  $p_j^k$  is the steady state probability of  $j$  customers in the M/M/1/k system. Using Knudsen (1972) and the fact that  $\left[ \frac{v\lambda}{\alpha} - \frac{h\lambda}{\alpha\mu} j \right]$  is monotone decreasing in  $j$ , we have that the function  $\Pi(\cdot)$  is unimodal. Hence we have the result. ■

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 above proposition implies that  $\hat{q}$  exists (with the possibility of being infinite).

Before we state the formal result, we need the following definitions.

**Definition 3.2** *We say that the threshold  $\bar{q}$  induces a pure strategy BNE if the pair  $(\tilde{g}(\cdot), \tilde{y}(\cdot))$  given by (3) forms a BNE, and this pair is said to be the induced BNE by this threshold.*

Since the focus of the paper is the ability of the firm to provide customers with unverifiable information, we will distinguish between two types of equilibrium in the cheap talk game (as well as in the delayed cheap talk game): an informative equilibrium and a babbling equilibrium.

**Definition 3.3** *We say that an equilibrium  $(y, g)$  is informative 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  $\sum_{\{q: g(q)=m_j\}} p_q(y, g) > 0$ .*

**Definition 3.4** *We say that a pure strategy MPBNE equilibrium strategy  $(y, g)$  is a pure strategy babbling equilibrium if  $y(m_i) = y(m_j)$  for all  $m_i, m_j \in \mathcal{M}$ .*

The following result characterizes all the equilibria of the above cheap talk game.

### Proposition 3.3

**Informative Equilibria:** *I. If  $q^* = \hat{q}$ , then  $q^*$  induces a pure strategy MPBNE.*

*II. If  $q^* > \hat{q}$ , there is no finite  $q$  that induces a pure strategy MPBNE.*

*III. If  $q^* < \hat{q}$ , then:*

*(a) If  $G(0, \hat{q}) > 0$ ,  $\hat{q}$  induces a pure strategy MPBNE.*

*(b) If  $G(0, \hat{q}) \leq 0$ , there is no finite  $q$  that induces a pure strategy MPBNE.*

**Babbling equilibria:** *There exists a pure strategy babbling equilibrium if and only if  $R \geq \frac{c}{\mu-\lambda}$ . Further, if  $q^* < \hat{q}$  and  $G(0, \hat{q}) < 0$ , i.e. Case III(b) above, there does not exist a pure strategy babbling equilibrium.*

To summarize the findings so far: we have identified three regions, each with a different equilibrium behavior. We observed that a pure strategy MPBNE exists only if the firm's and the customers' incentives are perfectly aligned or if the customers are mildly impatient. Proposition 3.3 establishes conditions for the existence of pure-strategy *informative* MPBNE's as a function of the system parameters and characterizes these whenever they exist. We also show the existence of a *babbling* equilibria, where the firm provides no information or the customers disregard any information the firm provides due to the firm's lack of credibility. Note that the firm may not admit a customer even if he himself 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.

In practice, the announcements made by the firm are rarely instantaneous. Using the above argument, one can observe that delaying the announcement can be advantageous in learning about the externality the arriving customer inflicts on other customers. The question is how does this delay impact the existence of equilibria in these service systems. We will first study the full access control of the firm, before analyzing the delayed cheap talk model.

#### 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 on current promotions (for example, Dominos Pizza announces the current deals) or inform customers regarding specific events, thus both the firm and the customer are obtaining some value, while incurring some waiting and holding costs in 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 an informative equilibria (and achieve its first best profit)?

We shall next define the model for the delayed announcements. We consider a system where the customers arrive according to a stochastic process. Each customer needs to be first processed by an automated

system (IVR) followed by the actual service (agent based service, ABS). The goal of the IVR in typical settings is 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.

*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  $\mu_{IVR}$ . 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 ABS which we model as a single server queue. We shall assume that the customer never reneges the system once he enters ABS.

*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  $R_{IVR} + R - c(w_{IVR} + w)$  where  $w_{IVR}$  is the waiting time in the IVR and  $W$  is the actual waiting time for the agent based service. We shall denote the random variable that corresponds to the sojourn time for the ABS by  $W$ . To define the utility function mathematically let  $I$  be the indicator function that represents the decision of the customer, i.e.,  $I = 0$  if the customer decides to balk the system and  $I = 1$  if the customer decides to join the system. Let  $W$  denote the virtual waiting time of the customer in the actual system. Utility of the customer can then be expressed as

$$U(I, w_{IVR}, W) = (R_{IVR} - cw_{IVR}) + (R - cW)I$$

Based on this it is clear that the customers balk only if  $R \geq c\mathbb{E}[W|S]$ , where  $\mathbb{E}[W|S]$  is the expected waiting time to the agents based service in equilibrium given the information provided by the system manager is  $S$ . (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 rigorously 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 service completion. At the same time, the firm also incurs an holding cost of  $h$  per customer per unit of time for any customer in the system (irrespective of the customer being in IVR or ABS). Let  $Q_{IVR}(t)$  and  $Q(t)$  denote that number of customers in IVR and ABS at time  $t$ , respectively. Let  $h$  be the holding cost of a customer for one unit of time, and  $D(t)$  is the counting process corresponding to the departure from the ABS. Then the firm's profit function is given by

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

where,  $\alpha$  is the discount factor. 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 instance at which the IVR completed 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 decides to balk the system after he is processed at the IVR only if  $cQ(t)/\mu > R$ . Note, that the customer treats the time before the announcement as “sunk cost,” and hence does not take it into account when deciding about the future. Note that this 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. Note that the threshold used by the customer is only based on the number of customers in ABS. Further, this threshold is identical to the one used by the customer in the base model where there was no delay in providing information. Hence, we shall denote this by  $\eta^c$ . Note that  $\eta^c = q^*$  defined in Section 3.

## 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 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. This set-up is similar to the access control problem studied in the literature. Ku and Jordan (2002) and Ku and Jordan (2003) study this setup with nodes which are modeled as loss systems. In our 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 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 formulate the system manager's problem as an MDP, and 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  $N$ . The  $N^{th}$  system is identical to the one defined before with the modification that the IVR has  $N$  servers and there is no waiting in the IVR, i.e., IVR is a pure loss system. For this system we can use the uniformization approach to obtain the following optimality equations:

$$U^N(q_{IVR}, q) = \frac{1}{\alpha + N\mu_{IVR} + \mu + \lambda} [-h(q_{IVR} + q)]$$

$$\begin{aligned}
& + \lambda U^N(q_{IVR} + \mathbb{I}_{q_{IVR} < N}, q) + \mu(U_{n-1}(q_{IVR}, q - \mathbb{I}\{q > 0\}) + v\mathbb{I}\{q > 0\}) \\
& + \mu_{IVR} \min\{q_{IVR}, N\} \max\{U^N(q_{IVR} - 1, q + 1), U^N(q_{IVR} - 1, q)\} + (N - q_{IVR})\mu_{IVR}U^N(q_{IVR}, q)
\end{aligned}$$

Next, we define a map  $\mathcal{L}$  on real valued functions on  $\mathbb{Z}^2$ , and a sequence  $U_n^N$  for  $n = 1, 2, \dots$ , where  $U_0^N \equiv 0$  and  $U_n^N = \mathcal{L}U_{n-1}^N$  is given by

$$\begin{aligned}
U_n^N(q_{IVR}, q) &= \frac{1}{\alpha + N\mu_{IVR} + \mu + \lambda} [-h(q_{IVR} + q) + \lambda U_{n-1}^N(q_{IVR} + \mathbb{I}_{q_{IVR} < N}, q) + \mu(U_{n-1}^N(q_{IVR}, q - \mathbb{I}_{q > 0}) + v\mathbb{I}_{q > 0}) \\
& + \mu_{IVR} \min\{q_{IVR}, N\} \max\{U_{n-1}^N(q_{IVR} - 1, q + 1), U_{n-1}^N(q_{IVR} - 1, q)\} + (N - q_{IVR})\mu_{IVR}U_{n-1}^N(q_{IVR}, q)]
\end{aligned}$$

It is clear that  $U^N$  is a fixed point of this mapping. Using the theory of the semi-Markov decision process, we also have that  $U_n^N \rightarrow U^N$  as  $n \rightarrow \infty$ . To show concavity of  $U^N$ , we shall show that  $U_n^N$  is concave and  $U_n^N(q_{IVR}, q) \leq U_n^N(q_{IVR}, q - 1) + v$ . The proof is by induction. The result holds for  $n = 0$  by definition as  $U_0^N \equiv 0$ . Assume that  $U_{n-1}^N$  is concave in  $q$  and  $U_{n-1}^N(q_{IVR}, q) \leq U_{n-1}^N(q_{IVR}, q - 1) + v$ , we shall show that  $U_n^N$  is concave and  $U_n^N(q_{IVR}, 1) \leq U_n^N(q_{IVR}, 0) + v$ . First, note that  $h(q_{IVR} + q)$  is linear, and  $\lambda U_{n-1}^N(q_{IVR} + \mathbb{I}_{q_{IVR} < N}, q)$  and  $(N - q_{IVR})\mu_{IVR}U_{n-1}^N(q_{IVR}, q)$  are both concave in  $q$ , as  $U_{n-1}^N$  is concave in  $q$ . Second, consider  $f_1(q_{IVR}, q) \equiv (U_{n-1}^N(q_{IVR}, q - \mathbb{I}_{q > 0}) + v\mathbb{I}_{q > 0})$ . Using the concavity and the fact  $U_{n-1}^N(q_{IVR}, 1) \leq U_{n-1}^N(q_{IVR}, 0) + v$ , it is easy to see that

$$2f_1(q_{IVR}, q + 1) \geq f_1(q_{IVR}, q) + f_1(q_{IVR}, q + 2), \quad \text{for all } q \geq 0.$$

Thus  $f_1$  is concave in  $q$ . Lastly, define  $f_2(q_{IVR}, q) \equiv \max\{U_{n-1}^N(q_{IVR} - 1, q + 1), U_{n-1}^N(q_{IVR} - 1, q)\}$ . We shall show that

$$f_2(q_{IVR}, q + 1) - f_2(q_{IVR}, q) \geq f_2(q_{IVR}, q + 2) - f_2(q_{IVR}, q + 1), \quad \text{for all } q \geq 0. \quad (6)$$

To this end, we shall consider the following four cases which are based on where the function  $U_{n-1}^N$  attains its maximum value.

**Case I:** Assume that  $U_{n-1}^N(q_{IVR} - 1, q + 3) \geq U_{n-1}^N(q_{IVR} - 1, q + 2)$ . Then using concavity  $U_{n-1}^N$  in  $q$  give us

$$0 \leq U_{n-1}^N(q_{IVR} - 1, q + 3) - U_{n-1}^N(q_{IVR} - 1, q + 2) \leq U_{n-1}^N(q_{IVR} - 1, q + 2) - U_{n-1}^N(q_{IVR} - 1, q + 1).$$

Combining the above with the definition of  $f_1(q_{IVR}, q)$ , we obtain the desired inequality (6).

**Case II:** Assume that  $U_{n-1}^N(q_{IVR} - 1, q + 3) < U_{n-1}^N(q_{IVR} - 1, q + 2) \geq U_{n-1}^N(q_{IVR} - 1, q + 1)$ . Then the right-hand-side of (6) is zero, where the left-hand-side of (6) is  $U_{n-1}^N(q_{IVR} - 1, q + 2) - U_{n-1}^N(q_{IVR} - 1, q + 1) \geq 0$ , since  $f_2(q_{IVR}, q) = \max\{U_{n-1}^N(q_{IVR} - 1, q + 1), U_{n-1}^N(q_{IVR} - 1, q)\} = U_{n-1}^N(q_{IVR} - 1, q + 1)$ . The last equality follows by noting that  $U_{n-1}^N$  is concave in  $q$ , and hence

$$U_{n-1}^N(q_{IVR} - 1, q + 1) - U_{n-1}^N(q_{IVR} - 1, q) \geq U_{n-1}^N(q_{IVR} - 1, q + 2) - U_{n-1}^N(q_{IVR} - 1, q + 1) \geq 0.$$

Thus, we obtain the desired inequality (6).

**Case III:** Assume that  $U_{n-1}(q_{IVR} - 1, q + 3) < U_{n-1}^N(q_{IVR} - 1, q + 2) < U_{n-1}^N(q_{IVR} - 1, q + 1) \geq U_{n-1}^N(q_{IVR} - 1, q)$ . Here, note that the Then the right-hand-side of (6) is negative, where the left-hand-side of (6) is zero. Thus, we obtain the desired inequality (6).

**Case IV:** Assume that  $U_{n-1}(q_{IVR} - 1, q + 3) < U_{n-1}^N(q_{IVR} - 1, q + 2) < U_{n-1}^N(q_{IVR} - 1, q + 1) < U_{n-1}^N(q_{IVR} - 1, q)$ . The proof then follows by the definition of  $f_2$  and concavity of  $U_{n-1}^N$  in  $q$ . Thus, we obtain the desired inequality (6).

Using the fact that  $U_{n-1}^N(q_{IVR}, q) \leq U_{n-1}^N(q_{IVR}, q - 1) + v$  and the definition of  $U_n^N = \mathcal{L}U_{n-1}^N$ , one easily obtains that  $U_n^N(q_{IVR}, q) \leq U_n^N(q_{IVR}, q - 1) + v$ . Thus, we establish the fact that  $U^N(q_{IVR}, q)$  is concave.

Next we will show that for large  $N$ , the system performance obtained, when modelling the IVR as  $M/M/N/N$  system is close to the one obtained when modeling it as an  $M/M/\infty$  system. We define two random variables:  $Q_{M/M/N/N}$  which represents the number of customers in an  $M/M/N/N$  queue in steady state with arrival rate  $\lambda$  and service  $\mu_{IVR}$ ; and  $Q_{M/M/\infty}$  which represents the number of customers in an  $M/M/\infty$  queue in steady state with arrival rate  $\lambda$  and service  $\mu_{IVR}$ . Let  $U(q_{IVR}, q)$  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

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

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

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

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

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

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

$$\|U^N - U\| \leq Ke^{-\beta N}.$$

Thus, one gets  $U^N \rightarrow U$  as  $N \rightarrow \infty$ . Thus, concavity of  $U^N$  in  $q$  results in concavity of  $U$  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, which depends on the number of customers occupied by the interactive voice response, such that, with  $q$  customers waiting for the ABS, a customer completing the IVR stage should be admitted to the ABS, only if  $q < \eta^*(q_{IVR})$ . Note that the situation where there is no IVR can be viewed as one where the threshold function is a constant. For the rest of the paper, we will assume the following:

**Assumption 4.1** *The threshold  $\eta^*$ , 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 Jr (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 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. 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 informative equilibria, while possibly achieving its first best profit.

## 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 systems 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 informative equilibrium. In the base model, when the announcements are made immediately upon the arrival of the customer, we have seen that informative equilibria exist only in certain regions, specifically, an informative equilibrium exists only in cases I and III(a) outlined in Proposition 3.3, but fail to exist in regions II and III(b). 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. Let  $\mathcal{M}$  be 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 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 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  $(y, g) \in \mathcal{G} \times \mathcal{Y}$  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  $m \in \mathcal{M}$ ,

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

2. Fixing  $y, g$  solves:

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

The above definition mirrors the one given for the base model, yet it is stated here for mathematical completeness, since the two games are different. We next characterize conditions under which pure strategy informative equilibria exist.

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

We shall show that the queuing dynamics observed under any informative BNE (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  $(Q_{IVR}, Q)$  in this system 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.

The firm clearly has no incentive to deviate from the Full Control solution. Thus, to ensure informative equilibrium 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 this satisfies condition (1) of Definition 5.1, we need to ensure the following two conditions hold:

$$\int_0^\infty \int_0^{\eta(x)} \left[ R - c \frac{q}{\mu} \right] p(q_{IVR}, q) dq_{IVR} dq \geq 0, \quad (7)$$

$$\int_0^\infty \int_{\eta(x)+1}^\infty \left[ R - c \frac{q}{\mu} \right] p(q_{IVR}, q) dq_{IVR} dq < 0. \quad (8)$$

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 (7)), and when getting the signal that prescribes "balk," i.e.,  $y(m) = 0$  (condition (8)). The above conditions can be restated as  $R \geq \frac{c}{\mu} \mathbb{E}[Q|Q < \eta^*(Q_{IVR})]$  and  $R < \frac{c}{\mu} \mathbb{E}[Q|Q > \eta^*(Q_{IVR})]$ . 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):

$$\underline{\eta}^c = \frac{\int_0^\infty \int_0^{\eta^*(x)} qp(q_{IVR}, q) dq_{IVR} dq}{\int_0^\infty \int_0^{\eta^*(x)} p(q_{IVR}, q) dq_{IVR} dq}, \quad \bar{\eta}^c = \frac{\int_0^\infty \int_{\eta^*(x)+1}^\infty qp(q_{IVR}, q) dq_{IVR} dq}{\int_0^\infty \int_{\eta^*(x)+1}^\infty p(q_{IVR}, q) dq_{IVR} dq}. \quad (9)$$

Note that the system dynamics under full control dictate that  $Q_{ABS}(t) \leq \sup_q \eta^*(q)$ . Thus, we have  $\bar{\eta}^c \leq \sup_q \eta^*(q)$ . The above results are summarized in the following theorem.

**Theorem 5.2** *The delayed announcement game has a pure strategy BNE if and only if*

$$\underline{\eta}^c \leq \frac{R\mu}{c} \leq \bar{\eta}^c.$$

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

Notice that the two thresholds  $\underline{\eta}^c$  and  $\bar{\eta}^c$  are the expected number of customers in ABS given that the firm wants the customers to join and balk the system, respectively. Informally speaking, the  $\underline{\eta}^c$  is the “average” of the area under the threshold function  $\eta^*(\cdot)$  weighted with respect to steady state probability of the number of people in ABS. Similarly, the  $\bar{\eta}^c$  defined analogously with respect to the area above  $\eta^*(\cdot)$ . 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  $\eta^*$  into the two relevant thresholds  $\underline{\eta}^c$  and  $\bar{\eta}^c$ .

The theorem implies that customers join the system as long as the reward to cost ratio 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. 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.

**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.*

**Proof:** 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 they receive one signal and balks when they receive 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. ■

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 informative pure strategy MPBNE for the dealyed cheap talk game. Thus the

condition in this theorem is necessary and sufficient for the existence of pure strategy informative MPBNE. Further, if a pure strategy informative BNE exists, it is unique (in terms of the outcomes) and the firm attains its first best under this equilibrium. As for base model, the delayed cheap talk game may have a babbling equilibrium. The condition for such an equilibria to exist is identical to that for the base model. With conditions for the existence of informative equilibrium 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 detract from the possibility of credibly communicating unverifiable information. That is, we would explore if the firm gains or loses the ability to communicate credible information when the information provision is delayed. From Theorem 5.2, we have  $\bar{\eta}^c \leq \sup_q \eta(q)$ . The following bounds on  $\sup_q \eta(q)$  and  $\hat{q}$  would be useful for this study.

**Proposition 6.1** *We have the following*

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

**Proof:** 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. ■

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 informative equilibrium. The customers' patience can be measured in terms of  $R\mu/c$ , the threshold beyond which they will not join had they had full information regarding the system status. In the case of the base model, if  $R\mu/c > \hat{q}$ , the firm cannot support an informative equilibrium, whereas for the delayed cheap talk game if  $R\mu/c > \bar{\eta}^c$  then the firm cannot support an informative equilibrium. The above results, show that both of these thresholds  $\hat{q}$  and  $\bar{\eta}^c$  are bounded above by  $v\mu/h$ . Thus, if  $R/c > v/h$  then there is no informative equilibrium 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 information is obtained in the IVR, delaying the announcement will not help the firm to credibly communicate delay information. Thus, for these parameters, the firm does not gain or lose any ability which it had without delaying.

However, if  $\hat{q} < R\mu/c < \bar{\eta}^c$ , 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  $q^* \in [\underline{\eta}^c, \bar{\eta}^c]$ . Here one can view  $q^*$  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. In order to facilitate the comparison, one can describe the results of the base model in a similar fashion. That is, a necessary and sufficient condition for the existence of equilibrium can be written as  $q^* \in [\underline{q}, \hat{q}]$ , where  $\underline{q}$  is the expected number of customers in the system when the firm achieves its first best, which is the expected number of customers in an  $M/M/1/\hat{q}$  queueing system with the arrival rate  $\lambda$  and service rate  $\mu$ .

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 informative equilibrium 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  $S = [\underline{q}, \hat{q}]$  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  $S_d$ .

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 1 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 and the contraction region are subsets of  $[0, v\mu/h]$ .*

The main implication of the above corollary is that there is 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.

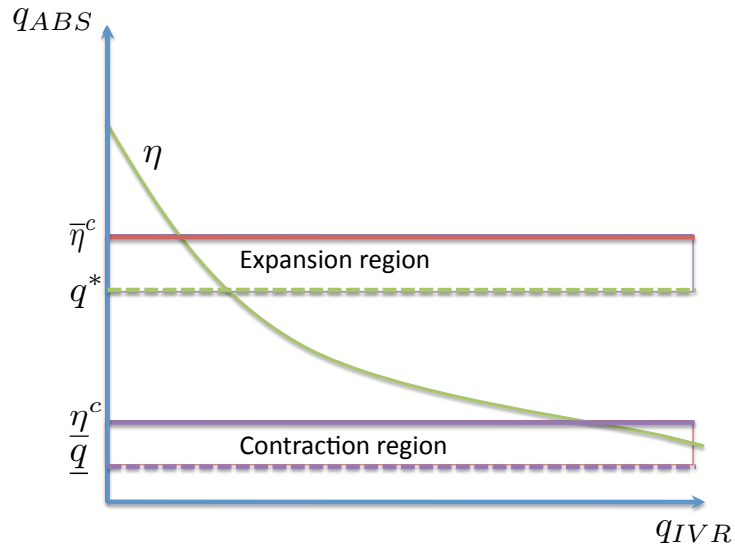
**Example 1:** For the first example, we assume the arrival rate,  $\lambda = 0.5$  customers per unit of time. The service rates in both the IVR and ABS are assumed to be unity, i.e.,  $\mu = \mu_{IVR} = 1$ . We assume that the firm obtains a value of  $v = 15$  from each served customer, yet incurs 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  $\eta^*$ , we compute the thresholds defined in (9)  $\bar{\eta}^c = 11.15$ ,  $\underline{\eta}^c = 1.99$ . Thus, using Theorem 5.2 we have that an informative equilibrium 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 informative equilibrium 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 or have no impact on the equilibrium language.

The existence of an expansion region above implies that we have a region in which an informative 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  $\eta^*$ , we compute the thresholds defined in (9)  $\bar{\eta}^c = 6.76$ ,  $\underline{\eta}^c = 3.58$ . Thus, using Theorem 5.2 we have that an informative equilibrium 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 informative equilibrium 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 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 1** 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 customer. 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. In order to accomplish this we assume that customers obtains 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  $R\mu/c$  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  $R\mu/c$  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.*

**Proof:**

(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  $p$  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:

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

The above implies

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

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

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

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. ■

The proposition shows that when the customers' threshold belongs to the expansion region, 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 informative equilibrium. Since the firm improves its profits going from the 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.

## 7. 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 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.

## 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. 2004. 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. 2007. The impact of delay announcements in many-server queues with abandonment. *Working Paper, Stern School of Business, NYU, NY* .
- Ata, B., T. Olsen. 2007. Dynamic leadtime quotation under general customer utilities. *Working paper, Kellogg School of Management, Northwestern University* .
- Crawford, V. P., J. Sobel. 1982. Strategic information transmission. *Econometrica* **50** 1431–1451.
- Dobson, G., J. Pinker. 2006. The value of sharing lead time information. *IIE Transactions* **38** 171–183.
- Duenyas, I., W. J. Hopp. 1995. Quoting customer lead times. *Management Science* **41**(1) 43–57.
- Ghoneim, H.A., S. Stidham Jr. 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.
- 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. *Working Paper* .
- Knudsen, N. C. 1972. Individual and social optimization in a multi-server queue with general cost-benefit structure. *Econometrica* **40** 515–528.
- 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.
- Spence, A. M. 1973. Job market signaling. *Quarterly Journal of Economics* **87** 355–374.