Received: 19 July 2021 | Accepted: 18 September 2022

DOI: 10.1111/poms.13903

ORIGINAL ARTICLE

Real-time delay announcement under competition

Siddharth Prakash Singh[1] | **Mohammad Delasay**[2] | **Alan Scheller-Wolf**[3]

[1]UCL School of Management, London, UK

[2]College of Business, Stony Brook University, Stony Brook, New York, USA

[3]Tepper School of Business, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA

Correspondence
Siddharth Prakash Singh, UCL School of Management, Level 38, One Canada Square, London E14 5AA, UK.
Email: siddharth.singh@ucl.ac.uk

Handling Editor: Michael Pinedo

Abstract

Internet-based technology enables firms to disseminate real-time delay information to delay-sensitive customers. We study how such delay announcements impact service providers in a competitive environment with two service providers who compete for market share. We model the service providers' strategies based on an endogenous timing game, investigating strategies that emerge in equilibrium. We determine the service providers' market shares under the various game outcomes by analyzing continuous-time Markov chains, which capture customers' joining decisions, and by developing a novel computational technique to analyze the intractable asymmetric Join-the-Shortest Queue system, providing bounds on the market shares. We find that only the lower capacity service provider announces its real-time delay under intermediate system loads and highly imbalanced capacities. However, for most parameter settings, the mere presence of a competitor induces both providers to announce delays in equilibrium, leaving customers better off on average. We relate our findings to the single-provider delay announcement literature by discussing the impact of competition on service providers, delay announcement technology firms, and customers.

KEYWORDS

delay announcement, endogenous timing game, Markov chain, service systems

1 | INTRODUCTION

Many service providers use delay information to manage congestion by influencing their customers' patronage decisions. Internet-based technology advancements have enabled customers to be informed about their delay at multiple service providers simultaneously—even before physically interacting with any of them—to decide which provider to patronize. For example, the HCA East Florida hospital system publishes real-time estimated delays at its emergency rooms (ERs) on its website (HCA East Florida, 2019). Similarly, the paid *ERtexting* service allows hospitals to text their expected delays to a central server that broadcasts the information to the community (Sadick, 2012). Many restaurants also employ Internet-based applications, such as *Yelp Waitlist*, to disseminate the expected time-to-seat to their potential diners (Yelp, 2022).

When a service provider functions in *isolation*, the extant literature has documented the advantages of delay announcements for both the service provider and customers (e.g., Whitt, 1999). Multi-service provider (or *network*) settings where *all* providers announce equally rich delay information have also received attention recently. Specifically, the implication of delay information on network synchronization (or coordination) has been studied in the context of ambulance diversion and ERs (Deo & Gurvich, 2011; Dong et al., 2019). However, competitive service providers may not broadcast the same types of information. For example, the Allegheny Health Network in Pittsburgh announces real-time delays for its urgent care centers (Rittmeyer, 2019), whereas the competing University of Pittsburgh Medical Center network does not currently provide such information. Similarly, not all restaurants are subscribed to Yelp Waitlist (Yelp, 2022). Therefore, it is not atypical for customers to make patronage decisions based on *heterogeneously rich* delay information from multiple service providers. In such cases, customers could turn to historical information in the absence of real-time information from a service provider (Brian, 2021; Perez, 2015).

Accepted by Michael Pinedo after three revisions.

Given such information heterogeneity in practice, it is important to study the motivation behind them—specifically, when service providers should or should not announce real-time delay in the presence of competition. Accordingly, we answer the following research questions:

1. What operational factors drive a service provider's decision on whether to announce real-time delay or not in the presence of competition? How?
2. In the presence of competition, how does delay announcement technology impact customers, service providers, and the technology providers? How do these insights (in a heterogeneous competitive setting) compare with those for a single provider?

We consider a network setting where delay-sensitive customers make patronage decisions based on (potentially) heterogeneously rich delay information from two service providers with comparable services, cost, and quality, but possibly different service capacities. These service providers, namely, *A* and *B*, compete for market share and are considering disseminating truthful[1] *real-time* delay information. Either of the service providers can initiate announcing real-time delay information, and once a service provider does so, the other service provider decides whether to respond by initiating their own delay announcements. We model the service providers' strategies using an *endogenous timing game*. We determine the game outcome in various information settings, leveraging the analysis of continuous-time Markov chain (CTMC) models representing the dynamics of customers' patronage decisions based on the available information. Our analysis contributes methodologically by presenting a novel procedure to compute the first nontrivial bounds on the arrival rates for asymmetric Join-the-Shortest Queue (JSQ) systems.

Our analysis reveals that both *A* and *B* typically find it favorable, in equilibrium, to announce real-time information. However, there are cases in which the *system load* and *relative service capacities* affect their decisions. In particular, when service capacities are highly imbalanced and the system load is intermediate, only the lower capacity service provider announces real-time delay, in equilibrium. Our findings under the competitive setting differ significantly from the extant literature about monopolistic settings. Specifically, in competitive settings, firms are more likely to announce their real-time delay, typically to the benefit of the lower capacity firm. (In a monopolistic setting, a firm would only announce delay when service capacity is low, or when service capacity is high but load is sufficiently low.) Further, from the technology providers' perspective, competitive firms are keener adopters compared to monopolists. This leads to customers being better off than in the absence of delay announcement technology. This differs from findings from the single-provider literature, where external intervention may be necessary to induce service providers to announce delay and therefore make customers better off (Hassin, 1986).

We complement our main results with three extensions to our base model. First, we extend to applications where customers patronize based on expected *sojourn time* rather than expected *delay*. Then, we extend to the case where announcements cost the service providers (for example, the service providers may pay a subscription fee to a third-party firm, like Yelp, for real-time delay announcement infrastructure). In this case, we find that the equilibrium delay announcement strategy is much more nuanced when the cost is moderate: the equilibrium outcome could involve none, one, or both of the providers announcing, and is highly sensitive to system load and the service providers' relative service rates. When the cost is high, neither of the providers announces delay in equilibrium. Finally, we extend our model to customers balking if their wait is too long. All our findings carry over to this setting.

2 | LITERATURE REVIEW

We first briefly review the literature about delay estimators, as one of our modeling choices pertains to delay estimations. Then, we review and highlight our contributions to two literature streams related to delay information provision in *single-service provider* and *multi-service provider* settings.

Delay estimators

We consider that service providers announce (if they decide to do so) queue length (QL) delay (*expected delay* ≈ *queue length* × *average service time*), which is commonly used in Markovian first-come-first-served (FCFS) systems because of its computational simplicity and accuracy (Ibrahim & Whitt, 2009a). Extensions are available for systems with customer abandonment, priority service, and time-varying arrivals (Ibrahim & Whitt, 2009b, 2011; Jouini et al., 2009, 2015). Ibrahim (2018) provides a comprehensive review of the delay announcement literature, including various types of delay estimators.

Delay information provision in single-service provider settings

The impact of delay information provision has been studied extensively in *single*-service provider settings. Whitt (1999) and Armony and Maglaras (2004), among others, document the benefits of QL announcements to improve wait times and system utilization when customers use them to make joining decisions. Armony et al. (2009) model the equilibrium joining behavior of utility-maximizing customers in a multi-server queueing environment. Akşin et al. (2016) show how customers react to long announced delays by abandoning service. Jouini et al. (2011) study the impact of announcing different percentiles of the waiting time distribution to control the system congestion.

In these single-service provider settings, it is not always optimal for a firm to disclose delay information. For example, Hassin (1986) shows that a revenue-maximizing service provider prefers to suppress her QL delay when her service rate is high and the system load is low. Dobson and Pinker (2006) show that a self-interested firm may hide lead time information when customers have a sufficiently low level of heterogeneity in their patience levels. Guo and Zipkin (2007) show that broadcasting more precise information could degrade system performance and customer experience under some waiting cost distributions; therefore, a self-interested firm may intentionally provide incomplete information (Allon et al., 2011; Guo et al., 2022). Dimitrakopoulos et al. (2021) study a firm that reveals and hides its QL in alternating periods. They show that the firm's profit generally improves if the durations of the revealing and hiding periods are appropriate.

In the papers mentioned above, customers' alternative to joining is to balk if their expected wait time is longer than their patience level. In contrast, we primarily focus on modeling competition as a service alternative by considering a *two*-service provider setting. This distinction produces fundamentally different results concerning when a firm should announce delay information. The main effect of competition is generally to induce both service providers to announce delay in equilibrium, leaving small regions of the parameter space where only the lower capacity service provider announces delay in equilibrium.

Delay information provision in multi-service provider (network) settings

We now discuss network settings (settings with multiple service providers). Ambulance diversion in ERs is closely related to the practice of delay announcement; ERs can request diversion of ambulances to other hospitals during overcrowding periods. Considering decentralized threshold diversion policies, Enders (2010) establishes the optimality condition for the "never divert" policy, and Deo and Gurvich (2011) establish the Pareto optimality of the equilibrium in which ERs are always on diversion. Do and Shunko (2015) propose a centralized threshold policy that is Pareto improving over the decentralized policy. He and Down (2009) and Ramirez-Nafarrate et al. (2014) show that delay announcements improve network synchronization and customer wait times even when only a small proportion of customers use the information. Pender et al. (2018) and Dong et al. (2019) study the impact of announcing moving average delays in a network setting and find that such announcements can cause the realized delays at the service providers to oscillate.

Unlike the above papers that do not consider the explicit question of whether a service provider should announce delay or not and whose primary focus is on *coordination* or *centralization*, we investigate the impact of delay announcements on competing service providers' *market shares*. We thus allow the service providers to choose whether or not to announce delay, and to have different delay announcement policies in equilibrium. The most relevant paper to our work is Hassin (1996), which models two gas stations with equal service rates on a highway where drivers only observe the nearer station's queue and infer the farther station's expected delay conditioned on the expected delay at the nearer station. Hassin concludes that the station with the observable queue always attracts more demand; thus, the server with the observable queue has an advantage over the server with the unobservable queue. Altman et al. (2004) extend this model to heterogeneous service rates and hypothesize that the emerging equilibrium is not always of threshold type (unlike in Hassin, 1996); they support this assertion using a mixture of numerical and analytical arguments. Hassin and Milo (2019) study a two-server setting with one observable server and one unobservable server, but both with no waiting room, and find that the welfare-maximizing equilibrium does not necessarily match the equilibrium arising from the customers' individual optimal decisions.

In Hassin (1996), Altman et al. (2004), and Hassin and Milo (2019), service providers do not have the choice of revealing or hiding their delay information; the congestion level of one service provider is always observable to customers, while the congestion level of the other service provider is always unobservable. In contrast, we allow the service providers to *choose* their best delay announcement strategy in an endogenous timing game. Furthermore, the mentioned papers assume customers are sophisticated enough to compute in *real-time* the conditional expected delay of the unobservable queue, given the state of the observable queue; this requires exact knowledge about the operating parameters of the service providers and complex equilibrium calculations. Under this setting, Altman et al. (2004) demonstrate that when the service providers have asymmetric service rates, the equilibrium policy may not even be threshold-type (where customers join the visible queue when its number of customers is fewer than a threshold and join the invisible queue otherwise). Rather than assuming that customers can compute such an equilibrium (especially given that the space of equilibria cannot be restricted to threshold-type), we model customers as inherently less sophisticated (akin to boundedly rational customers in the Economics literature): When only one service provider announces real-time delay, customers do not analytically infer the expected congestion level of the unobservable (nonannouncing) service provider; instead, they use periodically updated historical delay through published reports or online resources. Such historical information influences customers' decisions (Dong et al., 2019; Pender et al., 2018). Our work is the first to explicitly consider such a dynamic setting in which providers can choose to announce, which naturally leads to situations with one announcing and one nonannouncing firm.

3 | MODEL SETUP

Consider a system with two single-server service providers, A and B, with exponentially distributed service times with means $1/\mu^A$ and $1/\mu^B$. The service providers offer comparable services and compete for market share. Customers arrive according to a Poisson process with rate Λ. We normalize time so that $\mu^A = 1$, and we consider the system load $\rho = \Lambda/(1+\mu^B) < 1$ (for system stability).

In systems such as dine-in restaurants and ERs, customers are often concerned with and informed about the expected delay before their service starts (Dong et al., 2019; Richard, 2016). Accordingly, we consider that delay-sensitive customers patronize the service provider they expect to experience a shorter queue delay. (Indeed, a *longer service time* may or may not be preferable in such systems. In applications such as take-out restaurants, a model in which customers decide based on sojourn time could be more appropriate; we report our results for this case in Subsection 7.1). In our main model, customers have two service alternatives (service providers A and B) and always receive service from one of them, that is, they never balk.[2] In Subsection 7.3, we extend our main model by considering customers' balking behavior and show that our insights are robust to this case.

Under the status quo neither provider announces delay information. Therefore, A and B act as independent $M/M/1$ queues with status quo expected delays

$$D_0^A = \frac{\lambda_0^A}{1-\lambda_0^A} \quad \text{and} \quad D_0^B = \frac{\lambda_0^B}{\mu^B\left(\mu^B - \lambda_0^B\right)}, \tag{1}$$

where the status quo effective demand rates λ_0^A and λ_0^B to A and B ($\Lambda = \lambda_0^A + \lambda_0^B$) are determined endogenously such that $D_0^A = D_0^B$; that is, customers' delay-minimizing patronage decisions leads to equal expected delays. This corresponds to the *Wardrop equilibrium* (Hassin, 2016, p. 207) and the routing mechanism in the Hassin (1996, p. 623) model in which service providers' queues are unobservable. By setting $D_0^A = D_0^B$, A's status quo effective demand rate follows:

$$\lambda_0^A = \begin{cases} \dfrac{1+(\mu^B)^2 - \Lambda(1+\mu^B) - \sqrt{\left((\Lambda+1) - \Lambda\mu^B + (\mu^B)^2\right)^2 - 4\Lambda(1-\mu^B)}}{2(1-\mu^B)} & \mu^B \neq 1\,, \\ \dfrac{\Lambda}{2} & \mu^B = 1 \end{cases} \tag{2}$$

and B's status quo effective demand rate follows $\lambda_0^B = \Lambda - \lambda_0^A$. By replacing λ_0^A and λ_0^B in Equation (1), the status quo expected delays follows:

$$\begin{aligned} D_0^A &= D_0^B \\ &= \begin{cases} \dfrac{1+(\mu^B)^2+\Lambda\left(1-\mu^B\right) - \sqrt{\left((\Lambda+1)-\Lambda\mu^B+(\mu^B)^2\right)^2 - 4\Lambda(1-\mu^B)}}{2\mu^B\left(\Lambda - \left(1+\mu^B\right)\right)} & \mu^B \neq 1 \\ \dfrac{2}{2-\Lambda} & \mu^B = 1 \end{cases}. \end{aligned} \tag{3}$$

3.1 | The delay announcement game

Service providers A and B are considering initiating costless real-time QL delay announcements (we discuss the case of costly delay announcements in Subsection 7.2). Either provider may initiate announcing (thereby triggering a sequential game), or both may do so simultaneously (for example, when the technology simultaneously becomes available to them). To endogenize the timing of the decisions, we model them as an *endogenous timing game* (Hamilton & Slutsky, 1990). In this setup, A and B can decide whether to initiate delay announcements at the first opportunity or observe their competitor's action before doing so. We consider that the decision to initiate delay announcements (henceforth, *announce*) is irrevocable. In the endogenous timing game, time proceeds in two stages:

TABLE 1 Stage 1 game

	***B* announces**	***B* does not announce**
A announces	Regime $\mathcal{AB}$	Regime $\mathcal{A}$ or $\mathcal{AB}$
A does not announce	Regime $\mathcal{B}$ or $\mathcal{AB}$	Regime $\mathcal{N}$, $\mathcal{A}$, $\mathcal{B}$, or $\mathcal{AB}$

- Stage 1: The service providers decide whether to announce. If both announce, they continue to do so indefinitely.
- Stage 2: The nonannouncing service provider(s) can revisit their decision after observing their competitor's decision in Stage 1. If one announces in Stage 1, the other decides whether to respond by announcing in Stage 2. If neither announces in Stage 1, they can revisit their decisions in Stage 2 (Table 1).

We elaborate more about the game setup in Appendix B in the Supporting Information. To analyze the game, we demarcate four information regimes based on the service providers' eventual announcement choices (after Stage 2): (i) Regime $\mathcal{N}$ in which neither provider announces; (ii) Regime $\mathcal{A}$ in which only A announces; (iii) Regime $\mathcal{B}$ in which only B announces; and (iv) Regime $\mathcal{AB}$ in which both announce.

If A and B announce in Stage 1, the outcome is Regime $\mathcal{AB}$. If A (respectively, B) announces in Stage 1 and B (respectively, A) does not, the outcome is either Regime $\mathcal{A}$ or $\mathcal{AB}$ (respectively, Regime $\mathcal{B}$ or $\mathcal{AB}$), depending on B's (respectively, A's) response in Stage 2. If neither provider announces in Stage 1, the outcome could be any of the regimes, depending on A's (respectively, B's) preference between Regimes $\mathcal{A}$ and $\mathcal{N}$ and $\mathcal{AB}$ and $\mathcal{B}$ (respectively, Regimes $\mathcal{B}$ and $\mathcal{N}$ and $\mathcal{AB}$ and $\mathcal{A}$).

Let Λ_i^A and Λ_i^B denote the service providers' *long-run time-average demand rates* (henceforth, long-run demand rates) under Regime i. As the total arrival rate $\Lambda = \Lambda_i^A + \Lambda_i^B$ is fixed exogenously, an increase in a provider's market share (Λ^A/Λ for A and Λ^B/Λ for B) is equivalent to an increase in its long-run demand rate. So, breaking ties in favor of not announcing, a service provider S prefers announcing in Regime i to not announcing in Regime j if and only if $\Lambda_i^S > \Lambda_j^S$.[3] Given these preferences, in Proposition 1, we characterize conditions under which each regime emerges in equilibrium by considering the optimal Stage 2 responses to all possible Stage 1 decision pairs and factoring in these responses when evaluating Stage 1 decisions. All proofs appear in Appendix A in the Supporting Information.

Proposition 1. *The endogenous timing game either has no equilibrium or has exactly one equilibrium. Specifically, no regime emerges in equilibrium if and only if condition (4) or (5) holds. Otherwise, exactly one regime emerges in equilibrium; the conditions for each regime are specified in the proof in Appendix A.*1 *in the Supporting Information; see Equations (SI.*1)*–(SI.*4)*.*

$$\underbrace{\Lambda_{\mathcal{B}}^{A} \geq \Lambda_{\mathcal{AB}}^{A}}_{Condition\ 1} \wedge \underbrace{\Lambda_{\mathcal{N}}^{B} \geq \Lambda_{\mathcal{B}}^{B}}_{Condition\ 2} \wedge \underbrace{\Lambda_{\mathcal{A}}^{B} < \Lambda_{\mathcal{AB}}^{B}}_{Condition\ 3} \wedge \underbrace{\Lambda_{\mathcal{N}}^{A} < \Lambda_{\mathcal{A}}^{A}}_{Condition\ 4}, \quad or \tag{4}$$

$$\underbrace{\Lambda_{\mathcal{A}}^{B} \geq \Lambda_{\mathcal{AB}}^{B}}_{Condition\ 1} \wedge \underbrace{\Lambda_{\mathcal{N}}^{A} \geq \Lambda_{\mathcal{A}}^{A}}_{Condition\ 2} \wedge \underbrace{\Lambda_{\mathcal{B}}^{A} < \Lambda_{\mathcal{AB}}^{A}}_{Condition\ 3} \wedge \underbrace{\Lambda_{\mathcal{N}}^{B} < \Lambda_{\mathcal{B}}^{B}}_{Condition\ 4}. \tag{5}$$

When (4) holds, there is no equilibrium because:

- Regime $\mathcal{AB}$ is not an equilibrium as A would rather B be the sole announcer (Condition 1);
- Regime $\mathcal{A}$ is not an equilibrium as B would rather respond than let A be the sole announcer (Condition 3);
- Regime $\mathcal{B}$ is not an equilibrium because if B initiates announcements, A will not respond (Condition 1), and B prefers Regime $\mathcal{N}$ to $\mathcal{B}$ (Condition 2);
- Regime $\mathcal{N}$ is not an equilibrium as A prefers Regime $\mathcal{A}$ to $\mathcal{N}$ (Condition 4), and so would rather initiate.

Analogous explanations hold for (5). We note that we obtain (4)–(5) without imposing any structure on the values of Λ_i^A and $\Lambda_i^B, i \in \{\mathcal{N}, \mathcal{A}, \mathcal{B}, \mathcal{AB}\}$. Otherwise, (4)–(5) never hold. This is because Condition 1 rarely holds (that is, each provider generally prefers both providers announcing to only their competitor announcing); even when it holds, we find numerically that Condition 2 does not (that is, each provider prefers to be the sole announcer than for neither provider to announce). Accordingly, a unique regime always emerges in equilibrium. Our primary goal is to understand which delay information regime emerges in equilibrium.

3.2 | Patronage decisions and Markov chain models

This section details the customers' patronage decisions and the resulting CTMCs used to analyze the long-run demand rates.

Regime $\mathcal{N}$

Under Regime $\mathcal{N}$, the long-run demand rates are equal to the status quo rates, that is, $\Lambda_{\mathcal{N}}^i = \lambda_0^i, i \in \{A, B\}$, as derived in Equation (2).

Regimes $\mathcal{A}$ and $\mathcal{B}$

For brevity, we explain the model under Regime $\mathcal{A}$ (the model under Regime $\mathcal{B}$ is analogous). Provider A announces its expected delay as $d_n = n^A/\mu^A = n^A$ (as $\mu^A = 1$) when n^A is its current number of customers. As B does not announce under Regime $\mathcal{A}$, customers make patronage decisions based on B's historical average delay which is updated over time. That is, unlike the model employed by Hassin (1996) and Altman et al. (2004), we do not model customers as being able to infer B's expected delay conditioned on A's delay announcement. Instead, we let customers rely on the available historical average delay at B; this type of information has become more readily available in the recent past (e.g., Perez, 2015; Groeger, 2019).

To model B's available delay information, we consider the notion of *updating periods*: Customers' knowledge about the *expected delay* at B is updated at the end of each period, assuming that periods are long enough for the system stationarity. For example, published delays for some ERs are based on annual averages (Groeger, 2019). We index the updating periods by t, referring to the status quo as Period 0 ($t = 0$). We denote the single-period effective demand rate and expected delay at service provider i in Period t by λ_t^i and D_t^i, respectively.

Under Regime $\mathcal{A}$ and in any Period $t \geq 1$, customers decide based on the most recent historical delay at the (unobservable) service provider B and the announced real-time delay at A. Accordingly, they join A if $d_n \leq D_{t-1}^B$, and B, otherwise; that is, they break ties in favor of A. This induces a threshold structure for the arrivals: When $n^A < C_t = \lfloor D_{t-1}^B + 1 \rfloor$, where $\lfloor\ \rfloor$ is the floor function, the respective arrival rates to A and B are Λ and 0; otherwise, the respective arrival rates are 0 and Λ. Accordingly, the *arrival threshold* C_t in Period t is the current system size at A up to which A attracts arrivals. Figure 1a,b presents the CTMCs for the queueing dynamics of A and B under Regime $\mathcal{A}$ (respectively, Models A and B).

Model A is a birth-death process with states representing the current number of customers n^A at A. Model B's state space $\{(n^A, n^B) : n^A = 0, \ldots, C_t; n^B = 0, 1, \ldots\}$ tracks both n^A and n^B (the current number of customers at B) to determine B's effective demand rate. The transitions from a general state (n^A, n^B) in Model B follow:

- A service completion at A (respectively, B) when $n^A > 0$ (respectively,$n^B > 0$), with rate $\mu^A = 1$ (respectively,

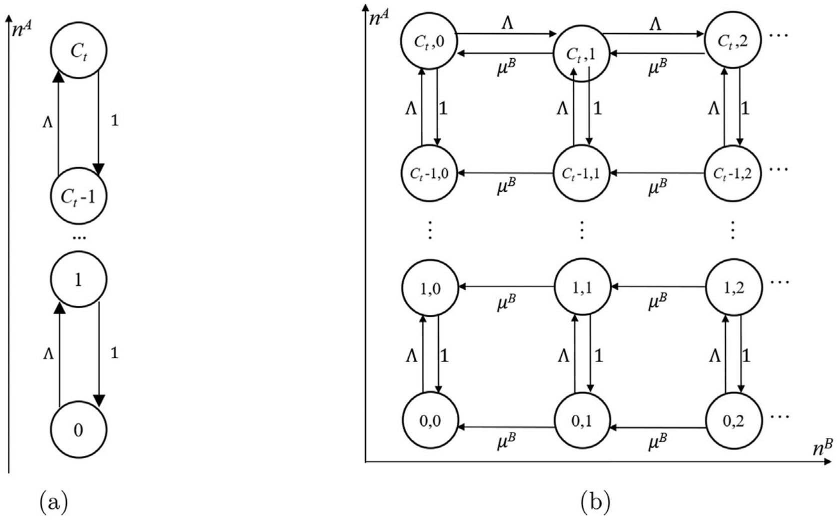


FIGURE 1 Regime $\mathcal{A}$ CTMCs in Period $t \geq 1$ ($\mu^A = 1$). (a) Model A for service provider A; (b) Model B for service provider B

μ^B), resulting in a transition to $(n^A - 1, n^B)$ (respectively, $(n^A, n^B - 1)$).

- An arrival to A (respectively, B) when $n^A < C_t$ (respectively, $n^A = C_t$), with rate Λ, resulting in a transition to $(n^A + 1, n^B)$ (respectively, $(C_t, n^B + 1)$).

In Period t, given the arrival threshold C_t, we analyze Model A to compute A's Period t effective demand rate λ_t^A, which yields $\lambda_t^B = \Lambda - \lambda_t^A$. Then, we analyze Model B to compute B's expected delay D_t^B, which determines the arrival threshold $C_{t+1} = \lfloor D_t^B + 1 \rfloor$, that is, the arrival threshold customers use to decide in Period $t+1$. Understanding the evolution of C_t over time is central to analyze Regime $\mathcal{A}$. Recall that we are interested in the long-run demand rate $\Lambda_{\mathcal{A}}^A$ (respectively, $\Lambda_{\mathcal{A}}^B$) under Regime $\mathcal{A}$; we compute this demand rate as the average effective demand rate to A (respectively, B) over T periods, where we let $T \to \infty$.

Regime $\mathcal{AB}$

The CTMC under Regime $\mathcal{AB}$ (Model AB shown in Figure 2) is a variant of the JSQ system, wherein a customer chooses the service provider with the shorter expected delay (breaking ties randomly) after comparing their QL delay estimates (i.e., $n^A/\mu^A = n^A$ vs. n^B/μ^B). The transitions from a general state (n^A, n^B) follow:

- A service completion at A (respectively, B) when $n^A > 0$ (respectively, $n^B > 0$), with rate $\mu^A = 1$ (respectively, μ^B), resulting in a transition to $(n^A - 1, n^B)$ (respectively, $(n^A, n^B - 1)$).
- An arrival to A (respectively, B) when $n^A < n^B/\mu^B$ (respectively, $n^A > n^B/\mu^B$), with rate Λ, resulting in a transition to $(n^A + 1, n^B)$ (respectively, $(n^A, n^B + 1)$). When $n^A \mu^B = n^B$, an arriving customer chooses a service provider randomly, resulting in transitions to $(n^A, n^B + 1)$ and $(n^A + 1, n^B)$, each at rate $\Lambda/2$.[4]

4 | ANALYZING LONG-RUN DEMAND RATES UNDER EACH INFORMATION REGIME

We analyze the long-run demand rates under $\mathcal{A}$ and $\mathcal{AB}$ in Subsections 4.1 and 4.2, respectively. Regime $\mathcal{N}$ is identical to the status quo and does not require additional analysis. Regime $\mathcal{B}$'s analysis is identical to Regime $\mathcal{A}$'s, with the indices transposed and time appropriately scaled.

4.1 | Analyzing Regime $\mathcal{A}$

To understand the long-run demand rates $\Lambda_{\mathcal{A}}^A$ and $\Lambda_{\mathcal{A}}^B$, we first evaluate the service providers' Period 1 demand rates under Regime $\mathcal{A}$ (i.e., their effective demand rate in the period immediately after A initiates real-time announcements solely), as they often help characterize long-run demand rates.

Period 1 *analysis*

Solving the balance equations of Model A (Figure 1a), we can derive its limiting probabilities $\pi_i, 0 \leq i \leq C_t$, in terms of the arrival threshold C_t and use them to derive A's effective demand rate in Period 1 as

$$\lambda_1^A = \Lambda \sum_{i=0}^{C_1 - 1} \pi_i = \frac{\Lambda^{C_1+1} - \Lambda}{\Lambda^{C_1+1} - 1}, \tag{6}$$

where D_0^B follows Equation (3) and $C_1 = \lfloor D_0^B + 1 \rfloor$. Using Equation (6), we can confirm that B, and hence the system, may become unstable in Period 1, as we characterize in Proposition 2.

Proposition 2. *Under Regime $\mathcal{A}$, the system could become unstable in Period* 1*, even if it is stable in Period* 0 (*i.e.,* $\rho < 1$). *Specifically, this occurs if and only if B's effective demand*

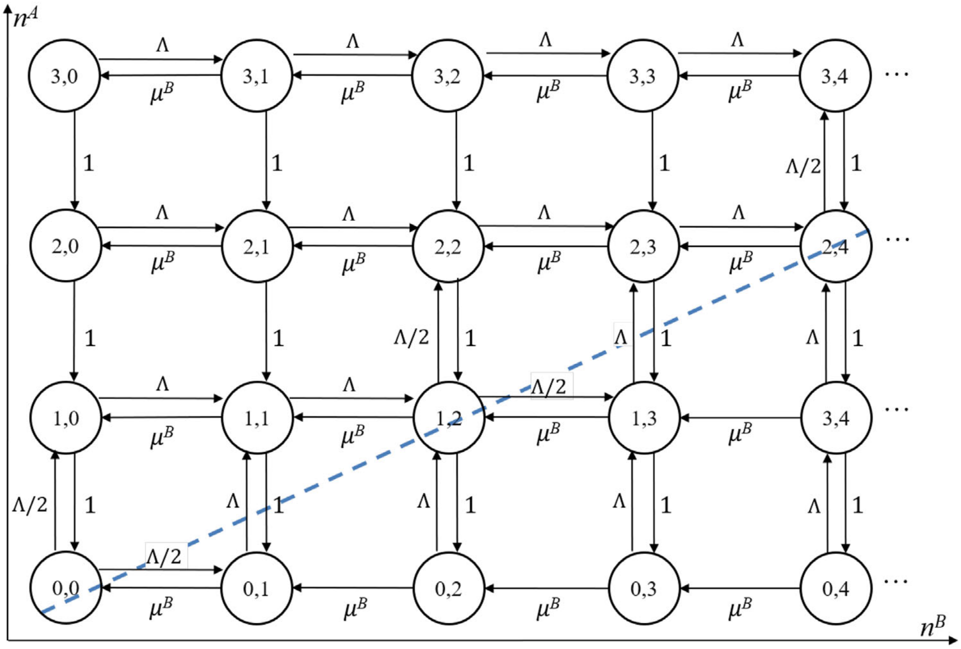


FIGURE 2 Regime $\mathcal{AB}$ CTMC (Model AB) in Period $t \geq 1$ ($\mu^A = 1$); the dashed line passes through states where tie-breaking is needed when $\mu^B = 2$ [Color figure can be viewed at wileyonlinelibrary.com]

rate in Period 1 *exceeds its service capacity* (*i.e.,* $\lambda_1^B \geq \mu^B$, *where* $\lambda_1^B = \Lambda - \lambda_1^A$).

We observe numerically (as explained in Section 6) that the system becomes unstable in Period 1 rarely; this may occur when B has significantly less service capacity than A. Moreover, if the system is stable in Period 1, it remains stable in all subsequent periods. All our remaining analysis in this section applies to the cases where the system remains stable in all periods.

Proposition 3 characterizes when announcing increases A's effective demand rate in Period 1. Under some conditions (presented in Proposition 4), these findings persist in the long term.

Proposition 3.

(a) *When A is the lower capacity service provider* (*i.e.,* $\mu^A < \mu^B$), *A's effective demand rate improves in the short term* (*i.e.,* $\lambda_0^A < \lambda_1^A$).
(b) *When A and B have equal service capacities* (*i.e.,* $\mu^A = \mu^B$), *A's effective demand rate either stays the same or improves* (*i.e.,* $\lambda_0^A \leq \lambda_1^A$), *depending on the system load* ρ.
(c) *When A is the higher capacity service provider* (*i.e.,* $\mu^A > \mu^B$), *A's effective demand rate may improve or worsen in the short term, depending on the system load* ρ.

Proposition 3 asserts that when A is the (weakly) lower capacity service provider, its Period 1 demand rate λ_1^A is (weakly) higher than her status quo demand rate λ_0^A. However, when A is the higher capacity service provider (i.e., $\mu^A > \mu^B$), as the example in Figure 3 shows, her Period 1 demand rate could be higher or lower than status quo, depending on the system load ρ. In this case, there are discontiguous regions of ρ for which λ_1^A is higher (these regions can be characterized by comparing Equations (2) and (6)).

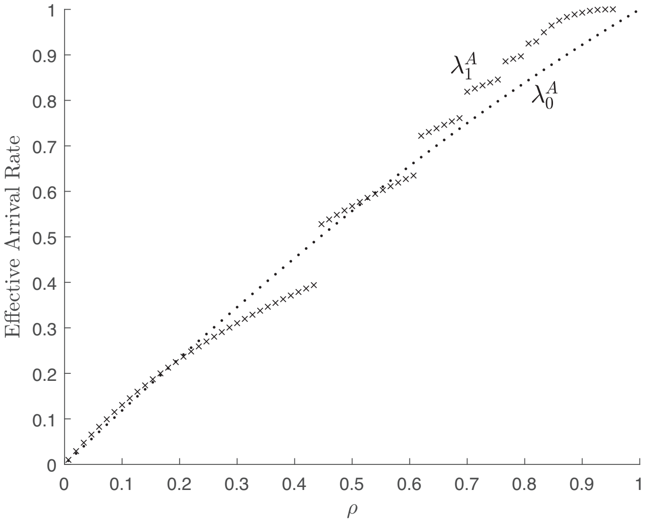


FIGURE 3 A's effective demand rates in Periods 0 and 1; $\mu^A = 1, \mu^B = 0.5$, $\Lambda \in (0, 1.5)$. The jumps in Period 1 rates correspond to the unit jumps in C_1 (and hence D_0^B) as ρ increases.

Long-term analysis

Regime $\mathcal{A}$'s Period 1 analysis allows us to analytically characterize the long-term dynamics when the arrival threshold remains the same in Periods 1 and 2 (i.e., $C_2 = C_1$). In

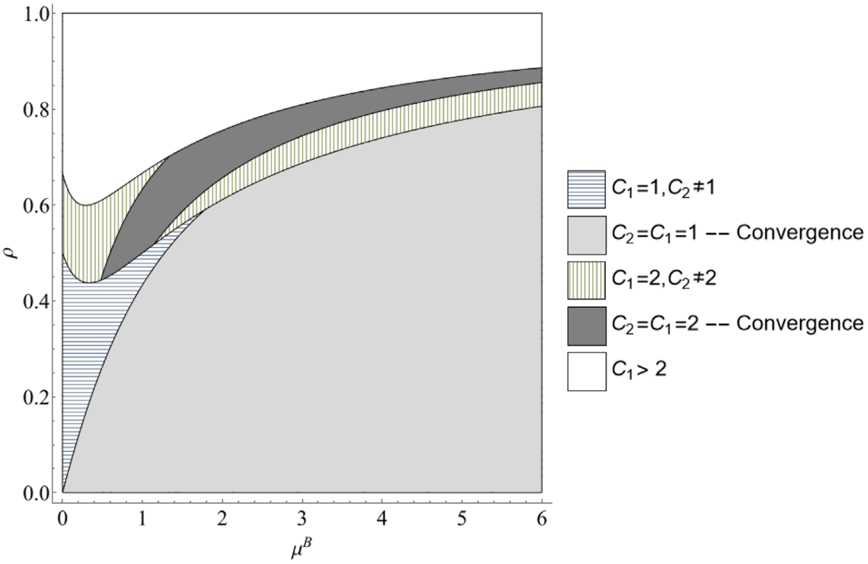


FIGURE 4 Solid shaded regions specify sufficient conditions for convergence in Period 2. Recall that $\mu^A = 1$. [Color figure can be viewed at wileyonlinelibrary.com]

this case, Periods 1 and 2 systems are identical, and therefore, the arrival threshold (and hence, the system) converges in Period 2 (i.e., $C_t = C_2 = C_1,\ \forall t > 2$). Consequently, the long-run demand rates still follow Equation (6), and therefore Proposition 3 holds in the long term. Proposition 4 provides analytical sufficient conditions for convergence in Period 2.

Proposition 4. *When the arrival threshold converges in Period* 2*, A's long-run demand rate under Regime $\mathcal{A}$ is identical to its Period* 1 *effective demand rate* (*as presented in Equation* (6)).

(*a*) *The arrival threshold converges to one in Period* 2 (*i.e.,* $C_1 = C_2 = 1$) *if:*

$$0 < \rho < \frac{1}{2(1+\mu^B)^2} \min\left\{ (\mu^B)^3 + \mu^B\sqrt{\left(\mu^B+5\right)\left(\mu^B+1\right)} - \mu^B, 2\left(\mu^B\right)^2 + \mu^B + 1 \right\}. \tag{7}$$

(*b*) *The arrival threshold converges to two in Period* 2 (*i.e.,* $C_1 = C_2 = 2$) *if analogously derived conditions to those in Part* (*a*) *of the proposition hold;* (*these conditions are unwieldy to present and can be downloaded from tinyurl.com/2p8a3e6w*).

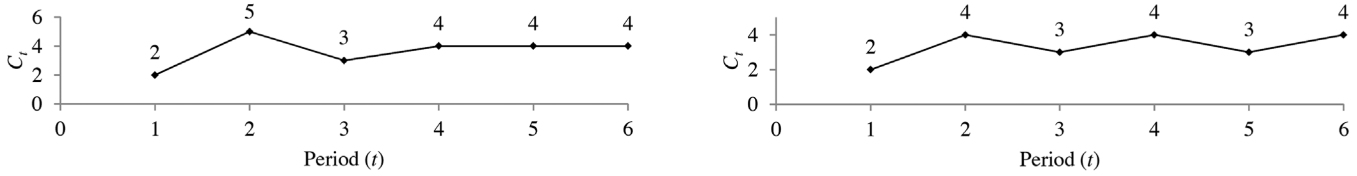
To obtain these conditions, we derive closed-form expressions for C_2 when $C_1 = 1$ or 2 by solving a simultaneous nonlinear system of equations. When $C_1 > 2$, this procedure results in higher order systems of equations, which in general cannot be solved in closed form, leading to the analytical intractability of Model B under a general arrival threshold C_t. Figure 4 specifies the parameter space (based on ρ and μ^B) where the conditions of Proposition 4 for $C_1 = C_2 = 1$ and $C_1 = C_2 = 2$ hold. Based on Proposition 4 and Figure 4, when the system load ρ is low to moderate, the long-term consequences of A's announcements are often identical to those in Period 1 (the period following initiating announcements).

For more general parameter settings, that is, when Proposition 4 does not hold, we need to understand the long-term behavior of the arrival threshold C_t. We define two possible long-term behaviors and prove in Proposition 5 that they are the *only* possible behaviors in the evolution of C_t:

- *Convergence*: The arrival threshold converges to a value, as in Figure 5a. This arises when $\exists t > 1 : C_t = C_{t-1}$, and therefore, $C_{t'} = C_t, \forall t' > t$. The conditions laid out in Proposition 4 are for a special case of convergence, wherein the arrival threshold converges in Period 2.
- *Stable oscillation*: The arrival threshold alternates between the same two values, as in Figure 5b. This arises when $\exists t > 2 : C_t = C_{t-2}$ and $C_t \neq C_{t-1}$, resulting in $C_{t'} = C_{t'-2}, \forall t' \geq t$

Proposition 5. *Under Regime $\mathcal{A}$ and when A and B are stable in all periods, the arrival threshold C_t either converges or establishes stable oscillation.*

If *convergence* occurs, the long-run demand rates are the rates associated with the arrival threshold to which the system converges. If *stable oscillation* occurs, the long-run demand rates are the average of the demand rates associated with the two arrival thresholds between which the system oscillates. Pender et al. (2018) and Dong et al. (2019) report similar oscillatory behavior, where the oscillation is due to the time lag between the reported delay and its effect on patronage decisions. The oscillations we observe are due to a similar lagged effect, where A's (and hence B's) expected delay in Period t is affected by B's expected delay in Period $t-1$. As an implication of Proposition 5, in the long term, the delay information used by customers in Period t is either (i) the realized expected delay in Period t (under convergence) or (ii) the expected delay that was realized in Period $t-1$ *and* that will be realized in Period $t+1$ (under stable oscillation).



(a)

(b)

FIGURE 5 Examples of stable long-term patterns of the arrival threshold C_t; in (a), C_t converges to 4 in Period 5, and in (b), it establishes a stable oscillation between 3 and 4 in Period 4. (a) Convergence; $\Lambda = 0.89, \mu^B = 0.5$; (b) Stable oscillation; $\Lambda = 0.87, \mu^B = 0.5$

So, this delay information will match the realized expected delay either in the immediate or the near future.

To fully characterize the evolution of the arrival threshold C_t under Regime $\mathcal{A}$ beyond Proposition 5, we need to explore: (i) when do convergence and stable oscillation occur? and (ii) at what value(s) of C_t does the system stabilize? Based on extensive numerical evidence, we conjecture the answers in Appendix C in the Supporting Information. Our conjecture implies that the customers' endogenous delay-minimizing behavior pushes the system into a steady state where the expected delay at B is near its lowest possible value, implying that C_t $(= \lfloor D_t^B + 1 \rfloor)$ stabilizes near its lowest possible value, and hence, the maximum QL (and hence, delay) at A is likewise near its lowest possible value. Utilizing this conjecture, we can obtain closed-form upper and lower bounds for A's long-run demand rates under Regime $\mathcal{A}$. We do not use these conjectured bounds further in this paper.

4.2 | Analyzing Regime $\mathcal{AB}$

Model AB (Figure 2) is similar to a JSQ system with asymmetric service rates, which is known to be analytically intractable (Adan et al., 1990). Some prior work presents provable bounds for stationary probabilities of the symmetric case (e.g., Halfin, 1985), but they do not extend to the asymmetric case. The extant literature on the asymmetric JSQ focuses on numerical approximations for the stationary probabilities and performance measures (e.g., Adan et al., 1990; Selen et al., 2016).

We contribute to this literature by proposing a computational algorithm that provides provable tight lower and upper bounds on the long-run demand rates under Regime $\mathcal{AB}$ when the system load ρ is low to moderate (up to 60-75%, depending on μ^B). (When ρ is high, our algorithm results in instability, and therefore, we do not obtain provable bounds.) The algorithm truncates Model AB's CTMC along its n^B dimension at some level T_B, and routes arrivals to A (regardless of n^A) when $n^B = T_B$. The truncation level T_B must be set to trade off appropriately between the tightness of the bounds and computational expense: A higher value of T_B provides tighter bounds, but they are more expensive to compute.

The resulting truncated CTMC is shown in Figure 6. The nonrepeating portion consists of states (i,j) such that $i \le \lceil j\frac{\mu^A}{\mu^B}\rceil \triangleq T_A$. T_A is the largest level at which customers arriving at $(T_A - 1, T_B)$ are routed to A. The repeating portion consists of states (i,j) such that $i > T_A$, in which the routing of arrivals does not depend on i: arrivals are routed to B (respectively, A) when $j < T_B$ (respectively, $j = T_B$). Given a starting state (i,j), we define the expected values R_j^A and R_j for the repeating portion and $N_{i,j}^A$ and $N_{i,j}$ for the nonrepeating portion as follows:

- R_j^A: The expected number of arrivals to A before n^A drops to $i-1$. (R_j^A is independent of i because of the chain's repeating structure, and so we do not need the index i in R_j^A.)
- R_j: The expected number of arrivals to the system before n^A drops to $i-1$.
- $N_{i,j}^A$: The expected number of arrivals to A before the next visit to state $(0,0)$.
- $N_{i,j}$: The expected number of arrivals to the system before the next visit to state $(0,0)$.

We use ideas similar to recursive-renewal-reward theory (Gandhi et al., 2014) to write recursive relationships between the above expected values, extending these ideas to yield provable bounds instead of exact values. We present our procedure in Algorithm 1, which uses systems of linear equations to compute upper bounds on R_j^A and $N_{i,j}^A$ (denoted by $\overline{R}_j^A$ and $\overline{N}_{i,j}^A$, respectively) and lower bounds on R_j and $N_{i,j}$ (denoted by $\underline{R}_j$ and $\underline{N}_{i,j}$, respectively). Using these computed bounds, we can compute an upper bound on $\Lambda^A_{\mathcal{AB}}$ (as we prove in Proposition 6).

Proposition 6. *When ρ is small to moderate (see the proof for a more precise characterization), $\Lambda^A_{\mathcal{AB}} \le \dfrac{1+\overline{N}^A_{0,1}+\overline{N}^A_{1,0}}{2+\underline{N}_{0,1}+\underline{N}_{1,0}}\Lambda$ where $\overline{N}^A_{0,1}, \overline{N}^A_{1,0}, \underline{N}_{0,1}$, and $\underline{N}_{1,0}$ are computed using Algorithm 1.*

Following the same procedure, it is trivial to compute a lower bound on $\Lambda^A_{\mathcal{AB}}$ by switching the roles of A and B and rescaling time appropriately. We use these bounds to determine

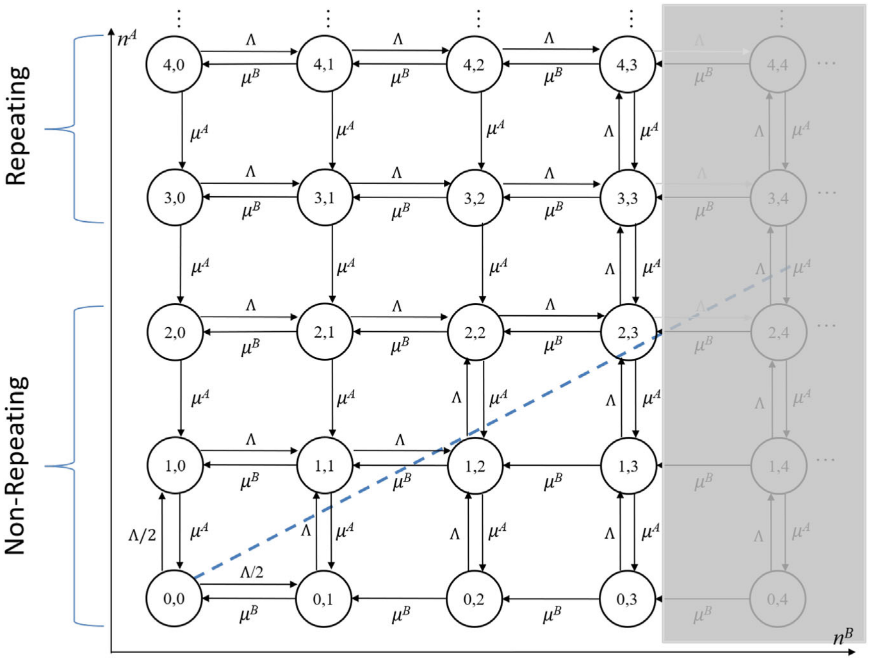


FIGURE 6 Truncated Markov chain for regime AB; $T_B = 3, \mu^B/\mu^A = 1.9$; grayed out portion is truncated [Color figure can be viewed at wileyonlinelibrary.com]

ALGORITHM 1 Procedure to compute an upper bound on $\Lambda^A_{\mathcal{AB}}$

```
Normalize time so that $\Lambda + \mu^A + \mu^B = 1$.
Set the truncation bound $T_B$. Define $T_A \triangleq \left\lceil j\frac{\mu^A}{\mu^B}\right\rceil$.
Solve the following system of $T_B + 1$ linear equations to obtain values of $\overline{R}^A_j$:
    $\overline{R}^A_j = \mu^B \overline{R}^A_{\max\{0,j-1\}} + \mathbb{1}_{\{j<T_B\}}\Lambda\overline{R}^A_{j+1} + \mathbb{1}_{\{j=T_B\}}\Lambda\left(1 + 2\overline{R}^A_{T_B}\right)$, for $j \in \{0, 1, \ldots T_B\}$.
Solve the following system of $T_B + 1$ linear equations to obtain values of $\underline{R}_j$:
    $\underline{R}_j = \mu^B \underline{R}_{\max\{0,j-1\}} + \mathbb{1}_{\{j<T_B\}}\Lambda(1 + \underline{R}_{j+1}) + \mathbb{1}_{\{j=T_B\}}\Lambda(1 + \underline{R}_{T_B} + \underline{R}_0)$, for $j \in \{0, 1, \ldots T_B\}$.
Solve the following system of $(T_A + 1)(T_B + 1)$ linear equations to obtain values of $\overline{N}^A_{i,j}$:
    $\overline{N}^A_{i,j} = \mu^B \overline{N}^A_{i,\max\{0,j-1\}} + \mu^A \overline{N}^A_{\max\{i-1,0\},j}$
        $+ \Lambda\mathbb{1}_{\{j/\mu^B < i/\mu^A \text{ and } j<T_B\}}\overline{N}^A_{i,j+1} + \Lambda\mathbb{1}_{\{j/\mu^B > i/\mu^A \text{ or } (j=T_B \text{ and } i<T_A)\}}\left(1 + \overline{N}^A_{i+1,j}\right)$
        $+ \Lambda\mathbb{1}_{\{i=T_A \text{ and } j=T_B\}}\left(1 + \overline{R}^A_{T_B} + \overline{N}^A_{T_A,T_B}\right)$, for $(i, j) \in \{0, 1, \ldots T_A\} \times \{0, 1, \ldots, T_B\}$.
Solve the following system of $(T_A + 1)(T_B + 1)$ linear equations to obtain values of $\underline{N}_{i,j}$:
    $\underline{N}_{i,j} = \mu^B \underline{N}_{i,\max\{0,j-1\}} + \mu^A \underline{N}_{\max\{i-1,0\},j}$
        $+ \Lambda\mathbb{1}_{\{j/\mu^B < i/\mu^A \text{ and } j<T_B\}}(1 + \underline{N}_{i,j+1}) + \Lambda\mathbb{1}_{\{j/\mu^B > i/\mu^A \text{ or } (j=T_B \text{ and } i<T_A)\}}(1 + \underline{N}_{i+1,j})$
        $+ \Lambda\mathbb{1}_{\{i=T_A \text{ and } j=T_B\}}\left(1 + \overline{R}^A_{T_B} + \underline{N}_{T_A,0}\right)$, for $(i, j) \in \{0, 1, \ldots T_A\} \times \{0, 1, \ldots, T_B\}$.
Compute $M = \frac{1 + \overline{N}^A_{0,1} + \overline{N}^A_{1,0}}{2 + \underline{N}_{0,1} + \underline{N}_{1,0}}$, which is an upper bound on the market share at $A$ under Regime $\mathcal{AB}$.
Recompute $\Lambda$ under the original time-scale (which was such that $\mu^A = 1$), and return $M\Lambda$.
```

the game outcome analytically in Section 5 and supplement our analysis with numerical experiments for a broader set of parameters in Section 6.

5 | ANALYTICAL DETERMINATION OF THE GAME OUTCOME

This section uses the long-run demand rates characterized in Section 4 to determine the equilibrium regime when A and B decide according to the endogenous timing game. We refer to these results as *analytical* since they either rely on long-run demand rates derived in closed form or on provable bounds (depending on the regime). Proposition 7 first characterizes the equilibrium outcome for the more tractable case of extreme system loads.

Proposition 7.

(*a*) *When the system load is sufficiently small* ($\rho \to 0$), *both providers announce delay in equilibrium* (*i.e., Regime* $\mathcal{AB}$ *emerges in equilibrium*).

(*b*) *When the system load is sufficiently large* ($\rho \to 1$), *neither provider announces delay in equilibrium* (*i.e., Regime* $\mathcal{N}$ *emerges in equilibrium*).

When the system load is very low, Regime $\mathcal{AB}$ emerges in equilibrium (Proposition 7(a)) because the lower capacity service provider has less than half the market share under Regime $\mathcal{N}$ but can reach a 50% market share by initiating delay announcements, which are then responded to by the other service provider (see Lemma 8(a) in Appendix A.7 for details in the Supporting Information). Proposition 7(b) confirms the intuitive result that since announcing delay has no impact on the long-run demand rates when the system load approaches 100% (see Lemma 8(b) in Appendix A.7 in the Supporting Information), neither A nor B has any incentive to initiate delay announcements.

For nonextreme system loads, we use results derived in previous sections, including Equation (2) for Regime $\mathcal{N}$, Equation (6) and Proposition 4 for Regime $\mathcal{A}$ (and their analogues for Regime $\mathcal{B}$), and the provable lower and upper bounds based on Algorithm 1 for Regime $\mathcal{AB}$, to establish which regime emerges in equilibrium (by substituting them into the conditions for each regime to emerge in equilibrium, given by Equations (SI.1)– (SI.4) in Appendix A.8 in the Supporting Information).

To cover a wide and reasonable parameter space, our numerical experiments consist of 43 values (as presented in Table 2) for the relative service capacities μ^B (recall that $\mu^A = 1$) between $1/6$ and 6 with more emphasis around $\mu^B = 1$. We also consider 200 equally spaced values of system load ρ in the range [1%, 97%] (we know from Proposition 7 that Regime $\mathcal{N}$ emerges in equilibrium when $\rho \to 1$).

Figure 7 presents the resulting equilibrium regimes when we can determine the game outcome *analytically* (about 33% of the whole parameter space).[5] According to the results, Regime $\mathcal{AB}$ generally emerges in equilibrium. Regimes $\mathcal{A}$ and $\mathcal{B}$ emerge in small regions when service rates differ significantly and the system load is intermediate.

TABLE 2 Subsets of different values of μ^B

$\mathbb{M}_{<<}$	$\{1/6, 1/5, 1/4, 2/7, 1/3, 3/8, 2/5, 3/7, 1/2, 4/7, 3/5, 5/8, 2/3, 5/7, 3/4, 4/5, 5/6, 6/7\}$
$\mathbb{M}_{<}$	{7/8,14/15,28/29}
$\mathbb{M}_{=}$	{1}
$\mathbb{M}_{>}$	{29/28,15/14,8/7}
$\mathbb{M}_{>>}$	{7/6,6/5,5/4,4/3,7/5,3/2,8/5,5/3,7/4,2,7/3,5/2,8/3, 3,7/2, 4, 5, 6}

The analytical limitations of Proposition 4 (which only yields closed-form solutions when we can analytically establish Period 1 convergence) and Algorithm 1 (which only converges for sufficiently small ρ and may yield loose bounds) do not allow us to determine the game outcome analytically for the remainder of the parameter space. Therefore, we determine the game outcome numerically in Section 6 and find that the above analytical indications continue to hold.

6 | NUMERICAL DETERMINATION OF THE GAME OUTCOME

For ease of exposition of determining the game outcome numerically for the remaining 67% of the parameter space, we denote the set of relative service capacities as $\mathbb{M}$ and partition it into five subsets as specified in Table 2. From the total of 8600 experiments, we remove eight experiments that lead to system instability (four occur under Regime $\mathcal{A}$ when $\mu^B = 1/6$, and four occur under Regime $\mathcal{B}$ when $\mu^B = 6$).

We employ matrix-analytic methods to compute the long-run demand rates under Regimes $\mathcal{A}$ and $\mathcal{B}$. For Regime $\mathcal{AB}$, we truncate n^A in Model AB at $\kappa = \lceil 10,000/\mu^B \rceil$ and n^B at $\lceil \mu^B \times \kappa \rceil$ to keep the truncation error negligible (the sum of stationary probabilities of the boundary states across all 8600 experiments is less than 0.000236). To reduce the effect of a tie-breaking rule for customers' patronage decisions, we increase each μ^B value in the sets $\mathbb{M}_{<<}, \mathbb{M}_{<}, \mathbb{M}_{>}$, and $\mathbb{M}_{>>}$ by a sufficiently small irrational ϵ. This favors B and removes the cases where the providers announce equal nonzero expected delays (recall that this occurs in states where $\mu^B n^A = n^B$). In state (0,0) for which ties are unavoidable, we break ties randomly with equal probability. Observe that we have experiments where the favored firm B has higher capacity (those in $\mathbb{M}_{>>}$ and $\mathbb{M}_{>}$) and where it has a lower capacity (those in $\mathbb{M}_{<<}$ and $\mathbb{M}_{<}$). We shall see that the equilibrium characterizations are qualitatively consistent in both situations, and therefore, that the tie-breaking rule does not have a significant impact on the outcome of the game.

Given our analytically determined outcome in Figure 7, we expect Regime $\mathcal{AB}$ to emerge in equilibrium for most parameter settings. Accordingly, we first examine when

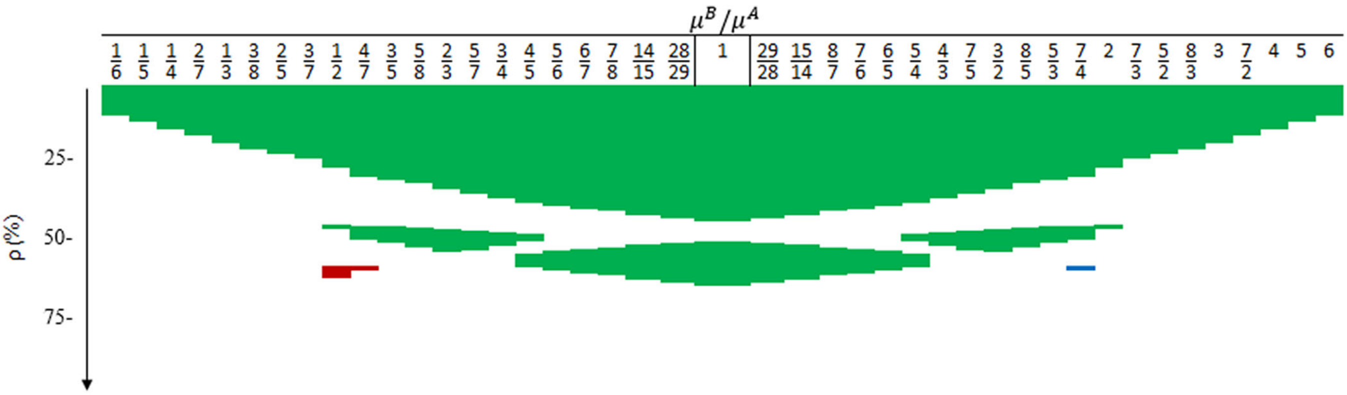


FIGURE 7 Analytically determined equilibrium regime; (green, yellow, blue, red, white) = ($\mathcal{AB}$, $\mathcal{N}$, $\mathcal{A}$, $\mathcal{B}$, no analytical determination) [Color figure can be viewed at wileyonlinelibrary.com]

Regime $\mathcal{AB}$ will emerge in equilibrium in Subsection 6.1. We treat all other cases in Subsection 6.2.

6.1 | When Regime $\mathcal{AB}$ emerges in equilibrium

Regime $\mathcal{AB}$ emerges in equilibrium when $\Lambda^B_{\mathcal{A}} < \Lambda^B_{\mathcal{AB}}$ and $\Lambda^A_{\mathcal{B}} < \Lambda^A_{\mathcal{AB}}$ (see condition (SI.4) in Appendix A in the Supporting Information), that is, when it is optimal for B to respond to A initiating announcements and vice-versa. Remark 1 summarizes our findings about when this occurs. We explain how our experiments lead to this finding after the remark.

Remark 1. A service provider almost always finds it favorable to respond to the competitor's announcement initiation unless it has a much higher service capacity and the system load is in an intermediate range. Therefore, Regime $\mathcal{AB}$ emerges in equilibrium unless one of the two service providers has a much higher service capacity and the system load is in an intermediate range.

We first explain the results when A initiates and B considers whether to respond. Across all 4996 experiments in $\mathbb{M}_{<<}$, $\mathbb{M}_{<}$, $\mathbb{M}_{=}$, and $\mathbb{M}_{>}$, B obtains a higher long-run demand rate by responding (i.e., $\Lambda^B_{\mathcal{AB}} > \Lambda^B_{\mathcal{A}}$); however, this holds for most, but not all, experiments (3533/3596) in $\mathbb{M}_{>>}$ with exceptions occurring when B's capacity is much higher ($\mu^B \geq 7/4$) and the system load ρ is intermediate. For example, in the shaded region in Figure 8, B does not respond when $0.75 \lesssim \rho \lesssim 0.79$. Figure 9 shows that this non-response region occurs for higher load values when service capacities are significantly imbalanced (i.e., when B's capacity is much larger than one). Similarly, when B initiates announcements, A finds it optimal to respond (i.e., $\Lambda^A_{\mathcal{B}} < \Lambda^A_{\mathcal{AB}}$) across all experiments in $\mathbb{M}_{>>}$, $\mathbb{M}_{>}$, $\mathbb{M}_{=}$, and $\mathbb{M}_{<}$. This is true for most, but not all, experiments (3495/3596) in $\mathbb{M}_{<<}$.

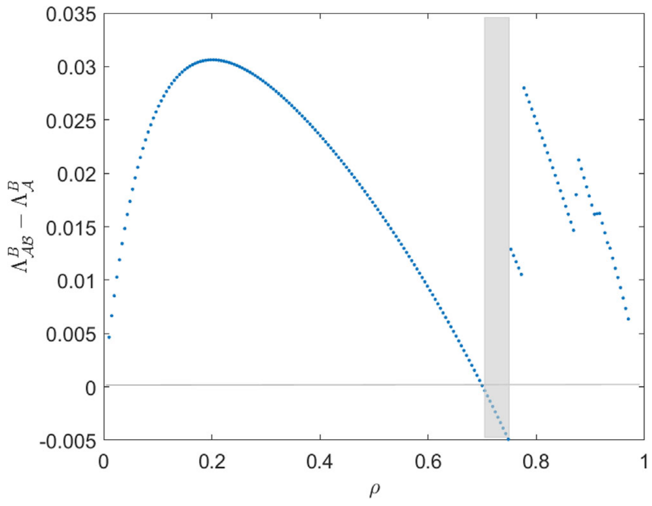


FIGURE 8 Changes in B's long-run demand rate when it responds; $\mu^B = 7/2$. B does not respond in the shaded load region. [Color figure can be viewed at wileyonlinelibrary.com]

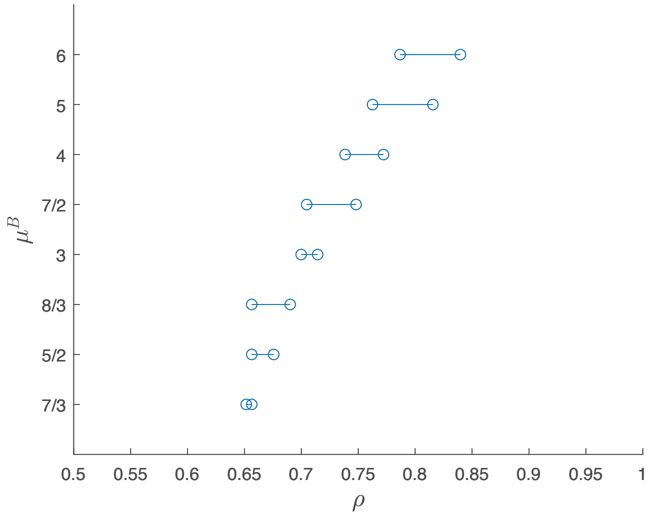


FIGURE 9 Ranges of intermediate system loads (for different values of μ^B) in which B does not respond to A. [Color figure can be viewed at wileyonlinelibrary.com]

The exceptions mentioned in Remark 1 arise because $\Lambda^B_{\mathcal{AB}} - \Lambda^B_{\mathcal{A}}$ depends on the system load ρ in a discontinuous and nonmonotonic fashion (as illustrated in Figure 8, for example). To keep our exposition in this section focused on the game outcome, we defer a more detailed explanation of the dependence of $\Lambda^B_{\mathcal{AB}} - \Lambda^B_{\mathcal{A}}$ on ρ to Appendix D in the Supporting Information. Remark 1 asserts that $\Lambda^B_{\mathcal{AB}} > \Lambda^B_{\mathcal{A}}$ except when (i) μ^B is much larger than μ^A and the system load is intermediate. Symmetrically Remark 1 also asserts that $\Lambda^A_{\mathcal{AB}} > \Lambda^A_{\mathcal{B}}$ except when (ii) μ^B is much smaller than μ^A and the system load is intermediate. From condition (SI.4) (in Appendix A in the Supporting Information), we have that Regime $\mathcal{AB}$ emerges in equilibrium unless (i) or (ii) holds. In Subsection 6.2, we focus on understanding what equilibrium emerges when (i) or (ii) holds.

6.2 | When Regime $\mathcal{AB}$ does not emerge in equilibrium

When Regime $\mathcal{AB}$ does not emerge in equilibrium, we have that either $\Lambda^A_{\mathcal{AB}} < \Lambda^A_{\mathcal{B}}$ or $\Lambda^B_{\mathcal{AB}} < \Lambda^B_{\mathcal{A}}$ (but not both; see Remark 1). Remark 2 summarizes our findings about when Regime $\mathcal{AB}$ does not emerge in equilibrium. We explain how our experiments lead to this finding after the remark.

Remark 2. A service provider prefers being the sole announcer to the situation with no announcers whenever their competitor does not find it favorable. In this case, the unique regime that emerges in equilibrium is the one where the initiator is the only announcer (i.e., Regime $\mathcal{A}$ or $\mathcal{B}$).

Based on whether $\Lambda^A_{\mathcal{AB}} < \Lambda^A_{\mathcal{B}}$ or $\Lambda^B_{\mathcal{AB}} < \Lambda^B_{\mathcal{A}}$, we deal with two cases:

Case 1: If $\Lambda^A_{\mathcal{AB}} \geq \Lambda^A_{\mathcal{B}}$ and $\Lambda^B_{\mathcal{AB}} < \Lambda^B_{\mathcal{A}}$, the possible equilibria and their associated conditions (obtained by simplifying conditions (SI.1)–(SI.4) in the Supporting Information) are:

- Regime $\mathcal{N}$: $\Lambda^A_{\mathcal{A}} \leq \Lambda^A_{\mathcal{N}} \leq \Lambda^A_{\mathcal{B}} \wedge (\Lambda^A_{\mathcal{AB}} = \Lambda^A_{\mathcal{B}} \vee \Lambda^A_{\mathcal{AB}} \geq \max\{\Lambda^A_{\mathcal{N}}, \Lambda^A_{\mathcal{B}}\})$.
- Regime $\mathcal{A}$: $\Lambda^A_{\mathcal{A}} > \Lambda^A_{\mathcal{N}}$.
- Regime $\mathcal{B}$: $\Lambda^B_{\mathcal{AB}} = \Lambda^B_{\mathcal{B}} \wedge \Lambda^B_{\mathcal{B}} > \Lambda^B_{\mathcal{N}}$.

For all 63 experiments for which $\Lambda^B_{\mathcal{AB}} < \Lambda^B_{\mathcal{A}}$, we have that $\Lambda^A_{\mathcal{A}} > \Lambda^A_{\mathcal{N}}$ and therefore, Regime $\mathcal{A}$ emerges in equilibrium. Furthermore, $\Lambda^B_{\mathcal{AB}} = \Lambda^B_{\mathcal{B}}$ never holds (indeed, $\Lambda^B_{\mathcal{AB}} < \Lambda^B_{\mathcal{B}}$ in all these experiments) and therefore, Regime $\mathcal{A}$ is the unique equilibrium.

Case 2: If $\Lambda^A_{\mathcal{AB}} < \Lambda^A_{\mathcal{B}}$ and $\Lambda^B_{\mathcal{AB}} \geq \Lambda^B_{\mathcal{A}}$, the possible equilibria and their associated conditions (obtained by simplifying conditions (SI.1)–(SI.4) in the Supporting Information) are:

- Regime $\mathcal{N}$: $\Lambda^B_{\mathcal{B}} \leq \Lambda^B_{\mathcal{N}} \leq \Lambda^B_{\mathcal{A}} \wedge (\Lambda^B_{\mathcal{AB}} = \Lambda^B_{\mathcal{A}} \vee \Lambda^B_{\mathcal{AB}} \geq \max\{\Lambda^B_{\mathcal{N}}, \Lambda^B_{\mathcal{A}}\})$.
- Regime $\mathcal{A}$: $\Lambda^A_{\mathcal{AB}} = \Lambda^A_{\mathcal{A}} \wedge \Lambda^A_{\mathcal{A}} > \Lambda^A_{\mathcal{N}}$.
- Regime $\mathcal{B}$: $\Lambda^B_{\mathcal{B}} > \Lambda^B_{\mathcal{N}}$.

For all 101 experiments for which $\Lambda^A_{\mathcal{AB}} < \Lambda^A_{\mathcal{B}}$, we have that $\Lambda^B_{\mathcal{B}} > \Lambda^B_{\mathcal{N}}$ and therefore, Regime $\mathcal{B}$ emerges in equilibrium. Furthermore, $\Lambda^A_{\mathcal{AB}} = \Lambda^A_{\mathcal{A}}$ never holds (indeed, $\Lambda^A_{\mathcal{AB}} < \Lambda^A_{\mathcal{A}}$ in all these experiments) and therefore, Regime $\mathcal{B}$ is the unique equilibrium.

6.3 | Game outcome

Putting together our observations from Remarks 1 and 2, we can assert the following outcome of the game: when the service providers have comparable or equal capacities, Regime $\mathcal{AB}$ emerges in equilibrium (Remark 1). When the service providers have significantly different capacities, Regime $\mathcal{AB}$ generally emerges in equilibrium, except for intermediate values of load, at which the equilibrium regime involves only the lower capacity service provider announcing delay (Remarks 1 and 2).

These findings are summarized in Figure 10. As expected, the equilibrium outcomes are near-symmetric about $\mu^B/\mu^A = 1$; the exceptions to symmetry are caused by our tie-breaking rule favoring B. The hatched cells in the figure represent parameters for which we had an analytically determined outcome in Figure 7; all the analytical determinations match the numerical outcome.

6.4 | Managerial insights

Using the equilibrium delay information regime characterization in Figure 10, we now state insights about how the availability of delay announcement technology affects various stakeholders in a setting with two competing firms. The stakeholders of interest are the service provider(s), the delay announcement technology firm, and the customers.

Effect on the service providers

The literature on announcing delay in monopolistic settings (see Dobson & Pinker, 2006; Hassin, 1986; Ibrahim, 2018) has established that a profit-maximizing service provider will announce delay (i.e., make its queue visible) when (1) its capacity is low or (2) its capacity is high and the system load is sufficiently low. In contrast, we find that two service providers under competition *almost always* announce delay, except that sometimes only the firm with significantly lower capacity announces at some intermediate load values. Nevertheless, the presence of competition makes the adoption of delay announcement technology almost inevitable because a single competitor announcing delay can generally capture a

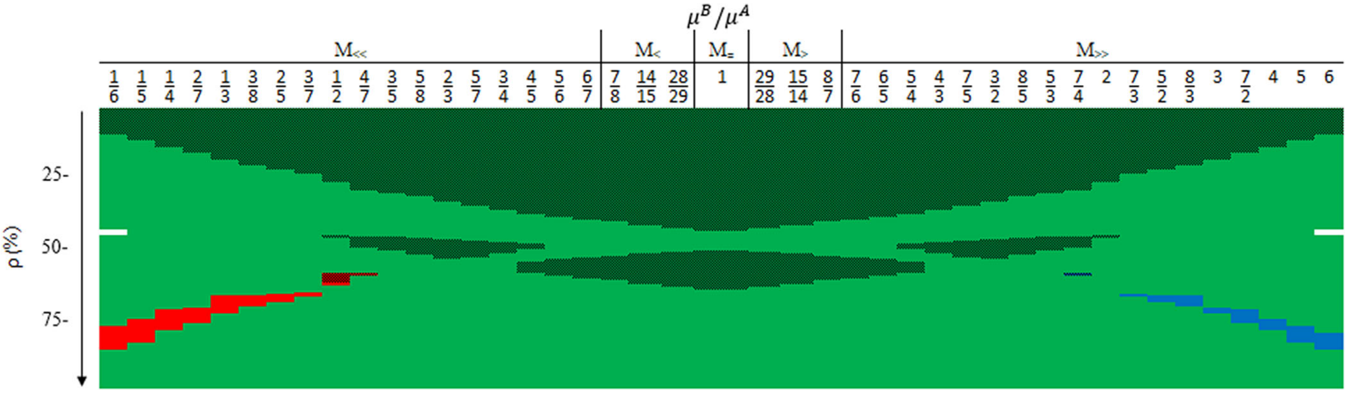


FIGURE 10 The equilibrium regime; (green, yellow, blue, red) = ($\mathcal{AB}$, $\mathcal{N}$, $\mathcal{A}$, $\mathcal{B}$). White cells represent instability. Hatched cells represent analytically determined outcomes. [Color figure can be viewed at wileyonlinelibrary.com]

large part of the market, prompting a competitive response. The resulting equilibrium market shares typically favor the lower capacity service provider.

Effect on the delay announcement technology firm

As both competing service providers almost always choose to announce delay in equilibrium, technology firms are likely to find keener adopters in a competitive setting. In equilibrium, the lower capacity service provider generally enjoys a larger market share than the status quo. Accordingly, the technology firm can benefit by marketing to the lower capacity service provider in a competitive setting by showing the projections of market share improvement. Therefore, we conclude that the competitor (the higher capacity service provider) will generally also be induced to adopt the technology. Our results imply that the technology firm should target market segments and geographies that feature competition rather than those that are monopolistic. To wit, the technology firm would be well-advised to target hospitals (or restaurants) in the vicinity of other similar hospitals (and restaurants).

Effect on customers

In a monopolistic setting, Hassin (1986) finds that, at intermediate system load values, external intervention is required to induce the service provider to announce delay to improve social welfare. Measuring customer welfare by the average delay they experience, we find numerically that the presence of delay announcement technology improves customer welfare for all our parameter settings (i.e., whether the equilibrium outcome is Regime $\mathcal{A}$, $\mathcal{B}$, or $\mathcal{AB}$). Thus, in a competitive environment, the presence of delay announcement technology improves customer welfare without the need for external intervention.

7 | MODEL EXTENSIONS

We investigate three extensions. In Subsection 7.1, we evaluate the outcome when customers use expected *sojourn time* instead of *delay* for patronage decisions. In Subsection 7.2, we evaluate the game outcome when the delay announcement technology firm charges a recurring cost (for example, a subscription fee). In Subsection 7.3, we evaluate the outcome when customers balk if they expect long delays.

7.1 | Patronage based on sojourn time

Our base model considers the more prevalent case where customers care more about their delay before their service starts (e.g., dine-in restaurants and ERs). There are also settings where customers likely care about sojourn time (delay before the service plus service time), such as take-out restaurants where customers value a short service time.

In the sojourn time setting, the equilibrium regime continues to be governed by Proposition 1. However, the long-run demand rates differ from those in Section 4. For instance, Regime $\mathcal{N}$'s analysis becomes more nuanced, as it may not be possible to split the total demand rate Λ to equalize expected sojourn times. To illustrate, let $\mu^A = 1$, $\mu^B = 6$, and $\Lambda = 1$; even if the entire market visits B, the resulting expected sojourn time $(1/(6-1) = 0.2)$ at B is less than the shortest possible expected sojourn time at A, which is 1. Therefore, in such situations, we assume that Λ would split so that B receives the entire market under Regime $\mathcal{N}$, although this does not equalize expected sojourn times. Similarly, Regimes $\mathcal{A}$ and $\mathcal{B}$ can have an arrival threshold $C_t = 0$. Therefore, many more parameter settings will result in instability at the nonannouncing provider.

Based on Figure 11, Regime $\mathcal{AB}$ continues to emerge in equilibrium for most of the parameter space, especially when the system load is relatively high. The sojourn time findings diverge most significantly from those of our original setting (Figure 10) in the following broad cases (in total, about 43% of our parameter space):

(i) When the system load is intermediate and service capacities are starkly different (the sojourn time model leads to instability).

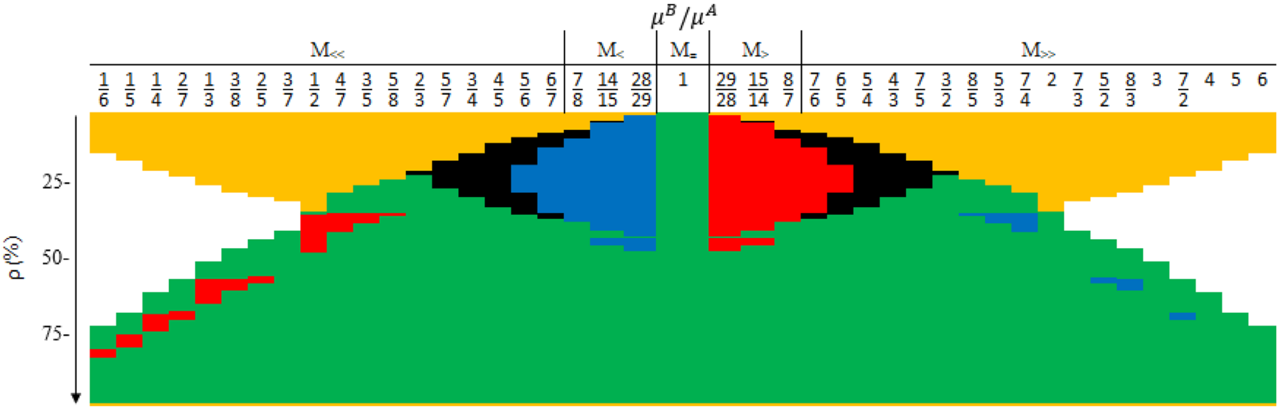


FIGURE 11 Numerically determined equilibrium regime when customers patronize based on sojourn times; (green, yellow, blue, red, black) = ($\mathcal{AB}$, $\mathcal{N}$, $\mathcal{A}$, $\mathcal{B}$, no equilibrium). White cells represent instability. [Color figure can be viewed at wileyonlinelibrary.com]

(ii) When the system load is low and service capacities are comparable (the sojourn time model may have no equilibria or may induce only the higher capacity service provider to announce delay).

(iii) When the system load is relatively low and service capacities are slightly different (the sojourn time model induces an equilibrium of Regime $\mathcal{N}$).

In the above cases, the patronage model based on delay induces an equilibrium regime of Regime $\mathcal{AB}$.

7.2 | Costly delay announcements

In our base model, delay announcements do not incur a recurring cost, for example, because the service providers have an in-house capability to announce delays. This section considers a recurring subscription cost incurred to employ the announcement technology. In this case, the announcement decision involves comparing the technology cost to the additional profit obtained through an increased demand rate. Considering a fixed profit $\nu > 0$ per customer served and a subscription cost $k \geq 0$ for the technology per unit of time (the same unit used to define μ^A, μ^B, and Λ), a service provider S would prefer announcing in Regime i to not announcing in Regime j if and only if $\nu\Lambda_i^S - k > \nu\Lambda_j^S \Leftrightarrow \Lambda_i^S - k/\nu > \Lambda_j^S$.

All our analytical results derived in Section 4, which characterize the long-run demand rates under each regime, remain valid under costly delay announcements as these results are independent of the subscription cost. Similar to Proposition 1 for the case of costless delay announcements, Proposition 8 characterizes the regime(s) that emerge in equilibrium under costly delay announcements.

Proposition 8.

(a) *No regime emerges in equilibrium if and only if the cost is such that:*

$$\max\{\Lambda_{\mathcal{AB}}^A - \Lambda_{\mathcal{B}}^A, \Lambda_{\mathcal{B}}^B - \Lambda_{\mathcal{N}}^B\} \leq \frac{k}{\nu} < \min\{\Lambda_{\mathcal{AB}}^B - \Lambda_{\mathcal{A}}^B,$$
$$\max\{\Lambda_{\mathcal{A}}^A - \Lambda_{\mathcal{N}}^A, \Lambda_{\mathcal{AB}}^A - \Lambda_{\mathcal{N}}^A\}\}, \textit{ or} \tag{8}$$

$$\max\{\Lambda_{\mathcal{AB}}^B - \Lambda_{\mathcal{A}}^B, \Lambda_{\mathcal{A}}^A - \Lambda_{\mathcal{N}}^A\} \leq \frac{k}{\nu} < \min\{\Lambda_{\mathcal{AB}}^A - \Lambda_{\mathcal{B}}^A,$$
$$\max\{\Lambda_{\mathcal{B}}^B - \Lambda_{\mathcal{N}}^B, \Lambda_{\mathcal{AB}}^B - \Lambda_{\mathcal{N}}^B\}\}. \tag{9}$$

(b) *Regimes $\mathcal{A}$ and $\mathcal{B}$ emerge in equilibrium when the cost is such that:*

$$\max\{\Lambda_{\mathcal{AB}}^A - \Lambda_{\mathcal{B}}^A, \Lambda_{\mathcal{AB}}^B - \Lambda_{\mathcal{A}}^B\} \leq \frac{k}{\nu}$$
$$< \min\{\max\{\Lambda_{\mathcal{A}}^A - \Lambda_{\mathcal{B}}^A, \Lambda_{\mathcal{A}}^A - \Lambda_{\mathcal{N}}^A\},$$
$$\max\{\Lambda_{\mathcal{B}}^B - \Lambda_{\mathcal{A}}^B, \Lambda_{\mathcal{B}}^B - \Lambda_{\mathcal{N}}^B\},$$
$$\times \max\{\Lambda_{\mathcal{A}}^A - \Lambda_{\mathcal{N}}^A, \Lambda_{\mathcal{B}}^B - \Lambda_{\mathcal{N}}^B\}\}. \tag{10}$$

(c) *Otherwise, one regime emerges in equilibrium, according to the conditions in Appendix A.8 in the Supporting Information.*

Proposition 8 implies that when delay announcements incur a cost, the game may result in no equilibria, one equilibrium, or multiple equilibria. Furthermore, we expect the equilibrium outcome to exhibit a much more complex relationship with ρ and μ^B compared to the costless case. As an illustration, observe in Figure 8 (by drawing a horizontal line at $\Lambda_{\mathcal{AB}}^B - \Lambda_{\mathcal{A}}^B = 0.025$ and separating points below and above the line) that when the cost-to-profit ratio $k/\nu = 0.025$ per time unit, B will find it worth the investment to respond to A's delay announcements when $0.18 \lesssim \rho \lesssim 0.40$ and $0.75 \lesssim \rho \lesssim 0.80$.

We evaluate the game outcome for different costs. To make a fair comparison across different relative service capacities (as profit is now proportional to load), we normalize the system capacity $\mu^A + \mu^B = 1$, maintaining the long-run demand rates $\Lambda_{\mathcal{N}}, \Lambda_{\mathcal{A}}, \Lambda_{\mathcal{B}}$, and $\Lambda_{\mathcal{AB}}$ on a consistent scale. (As an illustration, in order to fairly compare a setting with $\mu^A = 1$ and $\mu^B = 2$ to one with $\mu^A = 1$ and $\mu^B = 3$, we re-scale time in the first setting so that $\mu^A = 1/3$ and $\mu^B = 2/3$ and re-scale in the second so that $\mu^A = 1/4$ and $\mu^B = 3/4$; we use the re-scaled time to define the subscription cost k.)

For our experiments, we estimate a reasonable range for k/ν for the case of restaurants as an example. The average profit margin for a full-service restaurant is 3%–5% (Walters, 2019). Expanding this range to 1%-10% and assuming an average revenue per customer of \$6–\$50 (ProjectionHub, 2017), the profit range is ν of \$0.06–\$5 per customer. At capacity, a casual dining restaurant can serve on average 230 customers per day (ProjectionHub, 2017). Therefore, two restaurants together can serve 500 customers per day. As we have re-scaled time so that $\mu^A + \mu^B = 1$ and assuming the restaurants are open for 12 h a day, the normalized unit of time to use for k is $12 \times 60/500 = 1.44$ min. Monthly subscription costs for delay announcement applications are \$179–\$249 (Perez, 2017), leading to an estimated k of \$0.006–\$0.008 per time unit, assuming a 30-day month. Accordingly, a reasonable range of k/ν is 0.001–0.13. So, we run experiments with k/ν chosen from the set $\{0, 0.01, 0.05, 0.09, 0.13\}$. We have already discussed the results for $k/\nu = 0$ in the preceding sections, so we now turn our focus to discussing the remaining settings.

In Subsections 6.1–6.2, we observed and explained why when delay announcements are costless ($k/\nu = 0$), the equilibrium outcome is almost always Regime $\mathcal{AB}$. This remains the case when delay announcements are relatively inexpensive ($k/\nu = 0.01$). However, because of the nonmonotonicities and discontinuities as illustrated, for example, in Figure 8, a moderately large cost ($k/\nu = 0.05$) may complicate the equilibrium outcome. Indeed, in this case, there are also parameter settings with no equilibria, and with multiple equilibria.

On a high level, as expected, a higher cost dissuades the providers from announcing delay. In particular, as the cost increases, parameter settings that result in an equilibrium outcome of Regime $\mathcal{AB}$ shift to having an equilibrium outcome of Regime $\mathcal{A}$, $\mathcal{B}$, or $\mathcal{N}$. Similarly, as the cost increases, parameter settings that result in an equilibrium outcome of Regime $\mathcal{A}$ or Regime $\mathcal{B}$ shift to having an equilibrium outcome of Regime $\mathcal{N}$. If the equilibrium outcome is Regime $\mathcal{N}$ for a particular cost, it remains Regime $\mathcal{N}$ for higher costs. At the aggregate level, as the cost-to-profit ratio k/ν increases, the outcome shifts from Regime $\mathcal{AB}$ toward Regime $\mathcal{N}$ being dominant. Table 3 summarizes this trend.

At a more granular level, for a given cost and relative service capacity, the effect of increasing the system load has a complicated impact on the regime outcome. In general, increasing the load triggers multiple switches from one regime outcome to another. This complicated structure arises because of the nonmonotonic and discontinuous dependence of $\Lambda^A_{\mathcal{A}}$, $\Lambda^B_{\mathcal{A}}$, $\Lambda^A_{\mathcal{B}}$, and $\Lambda^B_{\mathcal{B}}$ on ρ. We also note that our analytically determined outcomes continue to predict the numerical outcome for a significant portion of the parameter space, ranging from 36% (when $k/\nu = 0.01$) to 56% (when $k/\nu = 0.05$). We provide further details in Appendix E in the Supporting Information.

7.3 | Incorporating customers' balking behavior

This section considers that customers have a finite, homogeneous patience level of $W_{\max}$. A customer compares the available delay information and routes to the service provider with the shorter expected delay, joining if the expected delay upon arrival is shorter than $W_{\max}$ and balking otherwise. Given the finiteness of $W_{\max}$, the system is guaranteed to be stable in all regimes. We now describe the associated modeling and analysis for each regime:

Regime $\mathcal{N}$

Under this regime, customers arrive at A and B at state-independent arrival rates and decide to join or balk upon arrival at the service provider: they balk at A (respectively, B) if $n^A \geq \lfloor W_{\max} + 1 \rfloor \triangleq C_A$ (respectively, $n^B \geq \lfloor (W_{\max}\mu^B) + 1 \rfloor \triangleq C_B$). Accordingly, A and B are $M/M/1/C_A$ and $M/M/1/C_B$ queues, respectively. In equilibrium, customers choose the respective arrival rates λ^A_0 and $\lambda^B_0 \equiv \Lambda - \lambda^A_0$ to A and B that equalize expected delays. Consequently, $\Lambda^A_{\mathcal{N}} = \lambda^A_0(1 - \Pr(n^A = C_A))$ and $\Lambda^B_{\mathcal{N}} = \lambda^B_0(1 - \Pr(n^B = C_B))$.

Regimes $\mathcal{A}$ and $\mathcal{B}$

Under Regime $\mathcal{A}$, customers join A if $n^A < \min\{\lfloor D^B_{t-1} + 1\rfloor, \lfloor W_{\max} + 1\rfloor\}$ and arrive at B otherwise; if they find $n^B/\mu^B > W_{\max}$, they balk from B. (As arriving customers balk from B if its expected delay is too long relative to $W_{\max}$, the average delay $D^B_{t'}$ at B in every period t' is smaller than $W_{\max}$. Therefore, the customer will always check at B if A's announced real-time delay is too long.) This system is amenable to exact numerical analysis because Model B (Figure 1b) is now finite along both dimensions. Regime $\mathcal{B}$'s structure follows similarly.

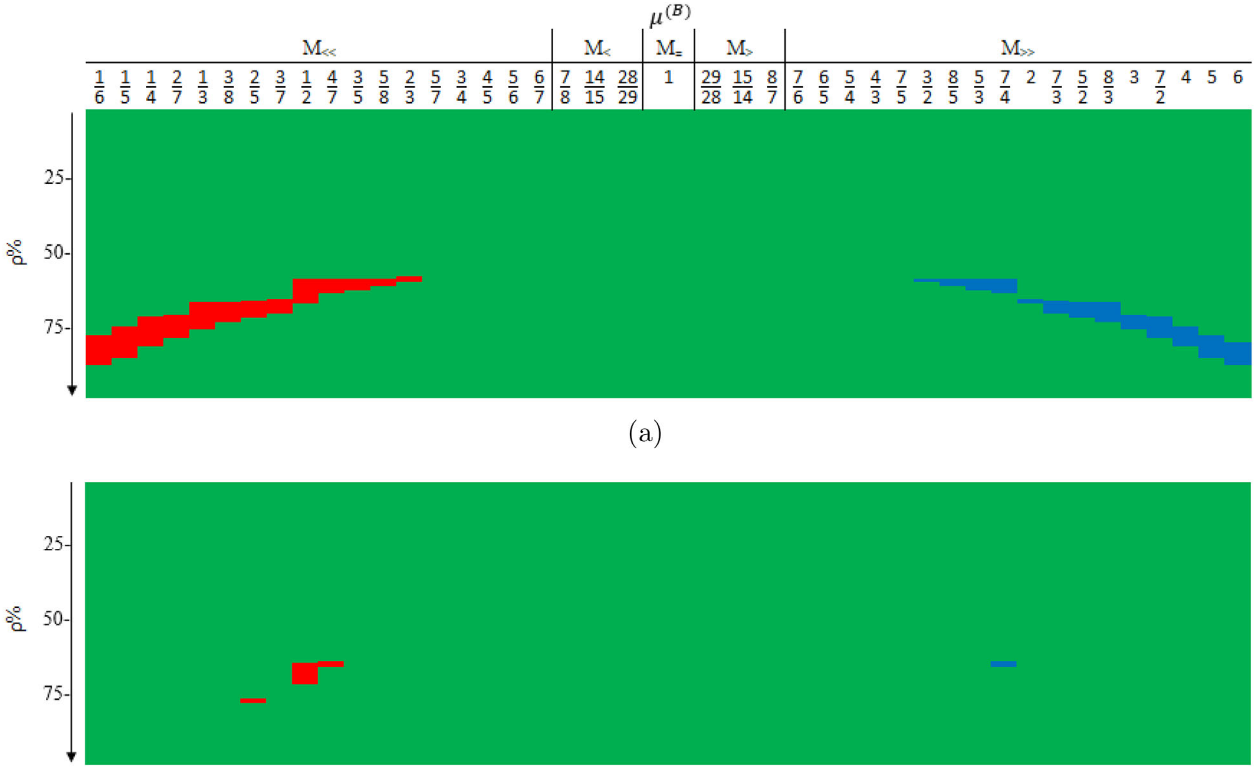
Regime $\mathcal{AB}$

Under this regime, customers join the service provider with the shorter real-time expected delay if it is shorter than $W_{\max}$. This results in a truncated version of the CTMC in Figure 2, based on which we can compute the long-run demand rates $\Lambda^A_{\mathcal{AB}}$ and $\Lambda^B_{\mathcal{AB}}$.

We report our results on the same set of parameters as in Section 6, normalizing system capacity $\mu^A + \mu^B = 1$. Recall from Subsection 7.2 that under this normalization, every unit of time corresponds to roughly 1.44 min. We consider $W_{\max} \in \{30, 10, 5\}$ (corresponding to patience levels of about 43, 14, and 7 min, respectively). Figure 12 summarizes the results. When $W_{\max}$ is high (Figure 12a),

TABLE 3 Percentage of experiments that result in each regime emerging in equilibrium

		Regime $\mathcal{N}$	Regime $\mathcal{A}$	Regime $\mathcal{B}$	Regime $\mathcal{A}$ and $\mathcal{B}$	Regime $\mathcal{AB}$	No equilibrium
Cost-to-profit ratio k/ν	**0**	0.00	0.73	1.18	0.00	98.09	0.00
	0.01	0.78	4.26	4.91	0.72	89.33	0.00
	0.05	23.89	19.12	19.43	15.14	18.34	4.07
	0.09	88.38	2.46	2.43	6.54	0.02	0.16
	0.13	100.00	0.00	0.00	0.00	0.00	0.00



(a)

(b)

FIGURE 12 Equilibrium regime when customers can balk; (green, yellow, blue, red) = ($\mathcal{AB}$, $\mathcal{N}$, $\mathcal{A}$, $\mathcal{B}$). (a) $W_{\max} = 30$, corresponding to a patience level of ~43 min; (b) $W_{\max} = 10$, corresponding to a patience level of ~14 min [Color figure can be viewed at wileyonlinelibrary.com]

we recover the same set of equilibrium regimes as the no-balking case (Figure 10) since customers rarely balk under any of the regimes. As $W_{\max}$ decreases (Figure 12b), Regime $\mathcal{A}$ or $\mathcal{B}$ emerge in equilibrium less frequently, with Regime $\mathcal{AB}$ taking their place. This is because the nonannouncer is more severely disadvantaged by the balking behavior; it receives overflow customers when the announcer is at its arrival threshold, but only until a threshold level (unlike in the no-balking case). So, when $W_{\max}$ is lower, $\Lambda^B_{\mathcal{A}} < \Lambda^B_{\mathcal{AB}}$ and $\Lambda^A_{\mathcal{B}} < \Lambda^A_{\mathcal{AB}}$ occur more frequently, leading to an equilibrium outcome of Regime $\mathcal{AB}$. As a result, when $W_{\max}$ is sufficiently low (among our experiments, when $W_{\max} = 5$), Regime $\mathcal{AB}$ always emerges in equilibrium.

8 | CONCLUDING REMARKS AND FUTURE DIRECTIONS

Technology advancements enable firms to disseminate delay information. In an endogenous timing game, we study whether a firm should initiate delay announcements when she competes for market share, uncovering the impact of relative service capacity and system load on the decisions in equilibrium. We find that only the lower capacity service provider announces its real-time delay under intermediate system loads and highly imbalanced capacities. However, for most parameter settings, the mere presence of a competitor induces both providers to announce delays in equilibrium.

Contribution to literature

Our results can be viewed as extending the results in Hassin (1986), Dobson and Pinker (2006), and Guo and Zipkin (2007) to a network setting. These papers establish when it may be suboptimal for a firm to reveal QL delay when the alternative to joining is to balk. In contrast, we explicitly model competition, that is, if a customer does not patronize A, she patronizes B. Furthermore, A and B's decisions are made endogenously in our model. This fundamental difference results in decisions that diverge from those in the above-mentioned papers. We can make the most direct comparison with Hassin (1986), in which customers are homogeneous (as in our model). Therefore, service capacity is the main driving force behind announcement decisions. Because Hassin (1986) models a single firm, the announcement decision dependence on service capacity is based on an *absolute* threshold: When the capacity is less than $2c/R$ (where R is the reward a customer obtains from completing service and c is the customer's waiting cost per unit time), it is always optimal for the service provider to reveal her queue; on the other hand, when the capacity is more than $2c/R$, the provider reveals her queue only when load is sufficiently low. In contrast, because we model two competing providers, the strategic interaction of these firms causes the initiation decisions for the two firms to violate such a threshold structure (see red and blue regions in Figure 10); this is a direct effect of the endogenous, rather than exogenous, outside option. Thus, the transition from a single service provider to a competing pair of providers who make announcement decisions fundamentally changes the system's dynamics. As discussed in Subsection 6.4, this difference has implications for firms (who are more likely to announce delay in the presence of competition), technology providers (who find that competing firms are keener adopters), and customers (who are always better off than in the status quo, without any external intervention).

Among the papers with two service providers, Hassin (1996) is the only one that studies the impact of delay announcements on market shares (When service rate are identical). Hassin finds that it is advantageous for one of the providers to reveal their queue, given that the other one does not. This corresponds to two of the regimes in our paper (Regime $\mathcal{A}$ and $\mathcal{B}$), for one particular parameter setting ($\mu^A = \mu^B$). Our paper generalizes the problem space by factoring in the possible response of the other service provider and enabling the two service providers to have *asymmetric* service rates. (Although Altman et al. (2004) consider asymmetric service rates, they do not factor in the possible response of the other provider, and their focus is on delays rather than market shares.) Methodologically, we contribute to the study of the asymmetric JSQ system by presenting Algorithm 1, which produces provable market share bounds for each of the service providers.

Future directions

There are numerous interesting avenues for future exploration. For instance, extending the analysis to include multi-server service providers, while analytically challenging, could lead to interesting results. Another potentially interesting extension is to model a heterogeneous customer population consisting of dedicated and flexible individuals: Dedicated customers would be loyal to one service provider regardless of her delay, while flexible customers would be delay-sensitive and patronize based on delay information. Such heterogeneity in the customer population is explored in He and Down (2009) and Dong et al. (2019) when the two service providers announce delay information of identical granularity. We expect that this extension would simply attenuate the effects of announcing. Furthermore, it may be interesting to investigate settings with three or more service providers. Finally, extending our bounding procedure for Regime $\mathcal{AB}$ to function for higher system loads would be of methodological interest as a further step toward characterizing the heretofore intractable asymmetric JSQ system.

ORCID

Siddharth Prakash Singh https://orcid.org/0000-0002-3337-547X
Mohammad Delasay https://orcid.org/0000-0001-9491-1136

ENDNOTES

[1] The backlash from systematic information misrepresentation could result in goodwill loss and legal actions; for example, O'Donnell (2014) describes the repercussions of delay information falsification at a Veterans' Administration Hospital.

[2] Given that customers choose the lower delay option, their patronage decision does not depend on their delay cost. Accordingly, the outcomes of our analysis are identical whether or not customers' delay costs are homogeneous.

[3] We shall see that in equilibrium A never finds it necessary to establish a preference between Regimes $\mathcal{A}$ and $\mathcal{AB}$ (respectively, Regimes $\mathcal{B}$ and $\mathcal{AB}$), so we can always break ties in favor of not announcing.

[4] We choose this tie-breaking rule because customers in our model are sensitive to delay; accordingly, equal delay announcements (with equal presumed accuracy) should result in equal demand rates at both providers.

[5] Note that our ability to analytically determine the outcome of the game is discontinuous in ρ. This is because we have analytical characterizations for Regime $\mathcal{A}$ and $\mathcal{B}$ for discontinuous regions of ρ; see Figure 4.

REFERENCES

Adan, I., Wessels, J., & Zijm, W. H. M. (1990). *Analysis of the asymmetric shortest queue problem*, Department of Mathematics and Computing Science, University of Technology.

Akşin, Z., Ata, B., Emadi, S. M., & Su, C.-L. (2016). Impact of delay announcements in call centers: An empirical approach. *Operations Research*, *65*(1), 242–265.

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

Altman, E., Jiménez, T., Núnez-Queija, R., & Yechiali, U. (2004). Optimal routing among ·/M/1 queues with partial information. *Stochastic Models*, *20*(2), 149–171.

Armony, M., & Maglaras, C. (2004). Contact centers with a call-back option and real-time delay information. *Operations Research*, *52*(4), 527–545.

Armony, M., Shimkin, N., & Whitt, W. (2009). The impact of delay announcements in many-server queues with abandonment. *Operations Research*, *57*(1), 66–81.

Brian, N. (2021). *Google announces skip the line: Restaurant wait times on search and maps*. https://restaurantclicks.com/google-announces-skip-the-line-restaurant-wait-times-on-search-and-maps/ (accessed August 30, 2022)

Deo, S., & Gurvich, I. (2011). Centralized vs. decentralized ambulance diversion: A network perspective. *Management Science*, *57*(7), 1300–1319.

Dimitrakopoulos, Y., Economou, A., & Leonardos, S. (2021). Strategic customer behavior in a queueing system with alternating information structure. *European Journal of Operational Research*, *291*(3), 1024–1040.

Do, H., & Shunko, M. (2015). Pareto-improving coordination policies in queueing systems with independent service agents. *Available at SSRN 2351965*.

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

Dong, J., Yom-Tov, E., & Yom-Tov, G. B. (2019). The impact of delay announcements on hospital network coordination and waiting times. *Management Science*, *65*(5), 1969–1994.

Enders, P. (2010). *Applications of stochastic and queueing models to operational decision making* (Ph.D. dissertation, Carnegie Mellon University).

Gandhi, A., Doroudi, S., Harchol-Balter, M., & Scheller-Wolf, A. (2014). Exact analysis of the M/M/k/setup class of Markov chains via recursive renewal reward. *Queueing Systems*, *77*(2), 177–209.

Groeger, L. (2019). *ER Inspector*. https://projects.propublica.org/emergency/ (accessed August 30, 2022)

Guo, P., Haviv, M., Luo, Z., & Wang, Y. (2022). Optimal queue length information disclosure when service quality is uncertain. *Production and Operations Management*, *31*(5), 1912–1927.

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

Halfin, S. (1985). The shortest queue problem. *Journal of Applied Probability*, *22*(4), 865–878.

Hamilton, J. H., & Slutsky, S. M. (1990). Endogenous timing in duopoly games: Stackelberg or cournot equilibria. *Games and Economic Behavior*, *2*(1), 29–46.

Hassin, R. (1986). Consumer information in markets with random product quality: The case of queues and balking. *Econometrica: Journal of the Econometric Society*, *54*(5), 1185–1195.

Hassin, R. (1996). On the advantage of being the first server. *Management Science*, *42*(4), 618–623.

Hassin, R. (2016). *Rational queueing*. CRC Press.

Hassin, R., & Milo, J. H. (2019). On rational behavior in a loss system with one observable queue and one unobservable queue. *International Conference on Queueing Theory and Network Applications* (pp. 166–182). Berlin: Springer.

HCA East Florida. (2019). *ER Wait Times*. https://hcaeastflorida.com/about/legal/er-wait-times.dot (accessed August 30, 2022)

He, Y.-T., & Down, D. G. (2009). On accommodating customer flexibility in service systems. *INFOR: Information Systems and Operational Research*, *47*(4), 289–295.

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

Ibrahim, R., & Whitt, W. (2009a). Real-time delay estimation based on delay history. *Manufacturing & Service Operations Management*, *11*(3), 397–415.

Ibrahim, R., & Whitt, W. (2009b). Real-time delay estimation in overloaded multiserver queues with abandonments. *Management Science*, *55*(10), 1729–1742.

Ibrahim, R., & Whitt, W. (2011). Wait-time predictors for customer service systems with time-varying demand and capacity. *Operations Research*, *59*(5), 1106–1118.

Jouini, O., Akşin, O. Z., Karaesmen, F., Aguir, M. S., & Dallery, Y. (2015). Call center delay announcement using a newsvendor-like performance criterion. *Production and Operations Management*, *24*(4), 587–604.

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

Jouini, O., Dallery, Y., & Akşin, Z. (2009). Queueing models for full-flexible multi-class call centers with real-time anticipated delays. *International Journal of Production Economics*, *120*(2), 389–399.

O'Donnell, K. (2014). *Head of troubled Phoenix VA hospital removed*. https://www.nbcnews.com/storyline/va-hospital-scandal/head-troubled-phoenix-va-hospital-removed-n255276 (accessed August 30, 2022)

Pender, J., Rand, R. H., & Wesson, E. (2018). An analysis of queues with delayed information and time-varying arrival rates. *Nonlinear Dynamics*, *91*(4), 2411–2427.

Perez, S. (2015). *Google search now shows you when local businesses are busiest*. https://techcrunch.com/2015/07/28/google-search-now-shows-you-when-local-businesses-are-busiest/ (accessed August 30, 2022)

Perez, S. (2017). *Yelp picks up restaurant waitlist app nowait for $40 million*. https://techcrunch.com/2017/03/01/yelp-picks-up-restaurant-waitlist-app-nowait-for-40-million/ (accessed August 30, 2022)

ProjectionHub. (2017). *4 financial projection models for the 4 restaurant styles*. https://blog.projectionhub.com/4-financial-projection-models-for-the-4-restaurant-styles/ (accessed August 30, 2022)

Ramirez-Nafarrate, A., Hafizoglu, A. B., Gel, E. S., & Fowler, J. W. (2014). Optimal control policies for ambulance diversion. *European Journal of Operational Research*, *236*(1), 298–312.

Richard, C. (2016). *Check wait times and get in line remotely with yelp*. https://blog.yelp.com/news/waitlist-times-get-line-remotely-yelp-nowait/ (accessed August 30, 2022)

Rittmeyer, B. C. (2019). *Allegheny health network releases wait-time tool for emergency rooms, urgent care centers*. https://triblive.com/local/pittsburgh-allegheny/allegheny-health-network-releases-wait-time-tool-for-emergency-rooms-urgent-care-centers/ (accessed August 30, 2022)

Sadick, B. (2012). *No wait at the ER*. http://health.usnews.com/health-news/articles/2012/09/12/no-wait-at-the-er (accessed August 30, 2022)

Selen, J., Adan, I., Kapodistria, S., & van Leeuwaarden, J. (2016). Steady-state analysis of shortest expected delay routing. *Queueing Systems*, *84*(3–4), 309–354.

Walters, S. (2019). *The average profit margin for a restaurant*. https://yourbusiness.azcentral.com/average-profit-margin-restaurant-13113.html (accessed August 30, 2022)

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

Yelp (2022). *How can I use Yelp Waitlist to get in line remotely*? https://www.yelp-support.com/article/How-can-I-use-Yelp-Waitlist-to-get-in-line-remotely?l=en_US (accessed August 30, 2022)

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Singh, S. P., Delasay, M., & Scheller-Wolf, A. (2022). Real-time delay announcement under competition. *Production and Operations Management*, 1–19. https://doi.org/10.1111/poms.13903