Self-interested customers populate various service systems. While self-interested, these customers may not be fully rational. We investigate the implications of boundedly rational customers on service networks where customers must visit multiple stations but can choose the order in which to visit the stations. We model customer strategy using anecdotal reasoning—that is, customers make decisions based on samples of system times experienced by customers who previously visited the system and followed each possible route. Using a fluid model, we fully characterize the evolution of customer routing decisions, with customers making decisions in each period based on anecdotal samples from the previous period. We completely specify the set of equilibrium routing profiles, where the fraction of customers choosing each route becomes stationary. In contrast with extant models of anecdotal reasoning, we find that anecdotally reasoning customers sometimes behave differently from fully rational customers, but not always. Equilibria can emerge under anecdotal reasoning that differ from fully rational equilibria, in which case, system performance suffers. Finally, we conduct extensive numerical tests, which demonstrate that our insights carry over to discrete systems that serve moderately many customers. If a service provider cannot control the service rates, then anecdotal reasoning can lead to waiting times that are much longer than necessary. However, although equilibria under customer anecdotal reasoning can differ significantly from those with fully rational customers, a service provider who controls the service rates would make the same decision under either form of customer reasoning, in both cases leading to a socially beneficial outcome.

Key words: open routing, bounded rationality, anecdotal reasoning, behavioral operations, queueing

This version: July 3, 2019.

1. Introduction

Self-interested customers populate various service systems, ranging from single-service systems like grocery stores and bank lines to multi-stage networks like the Department of Motor Vehicles (DMV) and airport check-in. This self-interest can manifest in myriad ways, and it is therefore important to
understand how customer behavior can affect system performance. Indeed, the rational behavior of self-interested customers has been studied in a wide range of service systems, with applications such as transportation networks (Braess 1968, Acemoglu et al. 2018); single-server queues (Naor 1969, Hu et al. 2017); queueing networks (Cohen and Kelly 1990, Honnappa and Jain 2015); restaurant reservations (Çil and Lariviere 2013); on-demand healthcare platforms (Liu et al. 2018); organ donation and transplantation (Su and Zenios 2004, Leshno 2017, Nageswaran and Scheller-Wolf 2017, Dai et al. 2018); and social networks and startups (Yang and Debo 2019).

In service systems with multiple stations, customers are often free to choose their routes through the network, creating an open-routing environment. In particular, in environments such as theme parks, shopping malls, and catered receptions, there are few structural restrictions on the sequence of stations that customers visit, and with this lack of structure comes greater freedom for customers to choose their own routes strategically. Practical case studies of service systems that fit the criteria for open routing, i.e., having multiple stations which need not be visited in a fixed sequence, are performed in Baron et al. (2016) and Shtrichman et al. (2001). Baron et al. (2016) study a medical clinic where multiple tests must be performed but the order is mostly irrelevant, and Shtrichman et al. (2001) discuss an army recruitment office where the recruits must submit to multiple independent evaluations. In these works the routing is flexible but centralized, so customers cannot self-select their routes. Systems with both open routing and self-interested customers have been studied by Parlaktürk and Kumar (2004) in a queueing model under steady state, and Arlotto et al. (2019) in a model where customers are present before the start of the service. In addition, Honnappa and Jain (2015) and Wang and Honnappa (2017) study a related setting called the “network concert queueing game,” which involves customers choosing their arrival times to a queueing network as well as their routes through the network. Thus far, the existing literature on strategic open routing has assumed the customers to be fully rational. However, for customers to arrive at a strategic equilibrium not only necessitates substantially more sophisticated reasoning than in simpler networks, but also requires them to be privy to—and keep track of—a non-trivial amount of information about the network (service rates, queue lengths, number of stations visited by other customers, etc.). Therefore, it is possible and perhaps even likely that customers will not behave like fully rational economic agents. Instead, they may adopt simple heuristics to decide on their routes.

The present work endeavors to understand the interaction between bounded rationality and open routing in a system where customers are present before the start of the service. We focus on a model with two stations in which customers must visit both stations, but they can freely choose the sequence of service. Customers want to minimize the total amount of time that they spend in
the system, but they are *boundedly rational* in that they rely on a simple and intuitive heuristic instead of computing the optimal solution under the assumption that all other customers are fully rational. The bounded rationality of customers is modeled using the classical *anecdotal reasoning* framework, discussed in Osborne and Rubinstein (1998) and Spiegler (2006). The word “anecdotal” conveys the idea that customers make their decisions based solely on anecdotes, i.e., one or more individual samples of outcomes (their own or someone else’s). Modeling self-interested customers as comparing anecdotal samples to make decisions constitutes a simple but surprisingly rich model of bounded rationality. This framework allows customers to operate without full knowledge of the model parameters or strong assumptions on the strategies of other customers, while still behaving in a manifestly self-interested fashion.

Under anecdotal reasoning, we allow the routing game to be played repeatedly, with new customers in each period who receive anecdotes drawn randomly from the outcomes in the previous period. To understand both the evolution and the equilibrium of customers’ behavior, we study a fluid approximation with a continuum of infinitesimal customers. In the fluid approximation, we fully characterize the evolution of customer routing decisions from one period to the next. This characterization in turn enables us to completely specify the set of equilibrium routing profiles, where the fraction of customers choosing each route reaches a stationary point. Furthermore, using numerical tests, we show that our fluid approximation accurately predicts outcomes for systems with at least a moderately large number of customers.

In many operational settings, researchers have observed outcomes under customer anecdotal reasoning that differ markedly from the outcomes with fully rational customers (see Section 2). In the open routing setting, however, we find that anecdotal reasoning can produce either similar or different outcomes when compared with the fully rational model, depending on the service rates at each station. In the fully rational model, Arlotto et al. (2019) observe that customers *herd* by all following the same route through the network; by contrast, in the model with anecdotal reasoning we observe that customers may or may not herd depending on the system parameters. Interestingly, customers under anecdotal reasoning herd when the ratio between the service rates for the slower and faster stations is either small or close to one, but they fail to herd when the ratio lies in a range between zero and one. Moreover, when customers do not herd, the distribution of customers taking different routes has a sharp transition at a particular threshold. This observation suggests that system performance under anecdotal reasoning is highly sensitive to the system parameters. Indeed, the average system time for customers can be as much as 35% higher at the anecdotal reasoning equilibrium than at the fully rational equilibrium. We derive managerial insights for service systems with open routing that allow a service provider to (i) understand the character of
strategic customer behavior in their systems under anecdotal reasoning, and (ii) use this knowledge to mitigate any performance degradation that could result from customer response to the design of the system.

The remainder of the paper is organized as follows. Section 2 reviews the related literature, and Section 3 describes the discrete model of customer routing decisions under anecdotal reasoning. We introduce the fluid approximation for customer routing and derive our main results in Section 4. Section 5 studies the performance of the system under customer anecdotal reasoning. In Section 6, we simulate a discrete version of our system to evaluate how closely the results hew to those from the fluid model. We discuss multi-sample versions of anecdotal reasoning in Section 7, and we conclude in Section 8. The proofs of all technical results can be found in the appendix.

2. Literature Review

There is an extensive literature on strategic customer behavior in service systems, beginning with Naor (1969). Surveys of this literature can be found in Hassin and Haviv (2003) and Hassin (2016). Recent work on strategic customer behavior in service systems includes Hassin and Roet-Green (2017), Yang and Debo (2019), Yang et al. (2019), and Cui et al. (2019), as well as several of the papers mentioned in Section 1.

There is also a burgeoning literature on modeling bounded rationality in operations management, which employs a variety of customer reasoning models such as quantal choice and logit choice (Su 2008, Chen et al. 2012, Huang et al. 2013, Li et al. 2016), among others. A survey of this literature can be found in Ren and Huang (2018). In particular, an area of work that has been actively incorporating bounded rationality is the study of strategic queueing: Li et al. (2016) on quality-speed competition; Hu et al. (2017) on “efficient ignorance,” where some but not all customers have queue state information; Wang and Hu (2017) on a setting where customers periodically share snapshots of the queue length with others deciding whether to join the system; Canyakmaz and Boyaci (2018) on rationally inattentive customers; Debo and Sniitkovsky (2018) on tipping and social norms; and Yang et al. (2018) on loss-averse customers.

We model customer bounded rationality through the concept of anecdotal reasoning, which was introduced in Osborne and Rubinstein (1998) as an alternative to the traditional Nash equilibrium concept traditionally used for noncooperative games. It is meant to more closely resemble human behavior by eschewing strong assumptions about players’ sophisticated reasoning and even their awareness of the game’s parameters, instead allowing customers to “sample” possible moves and learn from the outcomes. More recently, it has been studied by Spiegler (2006); in that work,
customers use $S(1)$-reasoning to learn about quality. The uncertainty is about the service quality of several providers, and customers sample from each provider once—hence the name $S(1)$-reasoning—and use that sample as their measure of the providers’ service quality.

Within the operations management literature, there are several recent papers that study customer anecdotal reasoning, e.g., Huang and Yu (2014) on opaque pricing, Huang et al. (2017) on posterior price matching, and Ren et al. (2018) on join-balk decisions in a queueing system. Importantly, much of this work focuses on customer inference about quality. By contrast, in our model customers make decisions about which route to follow through a service network, and to do so they must reason about the waiting time that they will face after choosing a given route.

To our knowledge, the only previous work to analyze customer anecdotal reasoning about waiting time in a queueing system is Huang and Chen (2015). They adopt the $S(1)$ reasoning framework, and customers make judgments about the expected waiting time in the queue based on a single sample. A customer joins if a waiting time equal to her sample would result in a positive net utility, given her unit waiting cost and the price of the service. They find that both the firm’s optimal prices and the equilibrium customer joining probability are both quantitatively and qualitatively different under anecdotal reasoning than with fully rational customers. Importantly, well-established pricing “rules of thumb” do not hold under anecdotal reasoning; the paper emphasizes the marked difference in outcomes between systems with fully rational customers and those with customers who employ anecdotal reasoning. The decision that customers face in our model is different from the join-balk decision in Huang and Chen (2015); in our context, customers cannot balk, but they must choose which of two stations to visit first. Moreover, we demonstrate the nuanced role that anecdotal reasoning plays in an open-routing service network, which distinguishes it from the results of Huang and Chen (2015) for a single-station system: depending on the model parameters, customers who use anecdotal reasoning may behave either exactly the same as or completely differently from fully rational customers.

Modeling bounded rationality is complex in almost any context. Our work constitutes a first attempt at studying the case of customer anecdotal reasoning about waiting time in the intricate confines of a service network, with multiple stations and in which customers must make routing decisions. In our context, the decision made by customers depends on the dynamics of more than a single queue. Applying the fluid approximation, we are able to fully characterize the evolution of customer routing decisions over time, as well as completely specify the space of equilibria. Moreover, in contrast to extant models of anecdotal reasoning, we identify conditions under which equilibrium routing decisions are the same under anecdotal reasoning as with fully rational customers, which has managerial implications for the capacity decisions of service providers.
3. The Model

We study a two-station service network with all customers present at the start of service. We label the stations as station A and station B. Figure 1 shows the network structure.

There are $N$ customers in the system. Customers can choose whether to visit station A first or station B first, and every customer must visit both stations. A customer who visits station A (B) first is said to have chosen route AB (BA). Service times at both stations are deterministic, and a service at station B is assumed to require strictly less time than a service at station A. Upon completing service at the first station on her chosen route, a customer immediately joins the back of the queue at the other station. When a customer makes her routing decision, she is not aware of the routing decisions made by others nor of the position that she will achieve in the queue at the first station on her chosen route. After customers make their routing decisions, customers who chose route AB are sequenced uniformly at random to determine the queueing order at station A, and similarly for BA customers—that is, those who chose route BA—at station B.

Our open-routing service network resembles that of Arlotto et al. (2019)—two stations, $N$ customers, customers must visit both stations. Importantly, they assume fully rational customers, using Nash equilibrium and subgame perfect equilibrium as solution concepts, while we study customers who reason anecdotally, as detailed below.

We suppose that our game is played repeatedly, by new players in each period $t \geq 1$. Let $x^{(t)}$ be the number of customers choosing route AB in period $t$. We use the term “system time” to refer to the time elapsed from when a customer enters the system to when she leaves the system. For a given customer, this quantity can thus be calculated as the sum for that customer of her times in queue at both stations and her service times at both stations. Let $A_i^{(t)}$, $i = 1, \ldots, x^{(t)}$, be the system time in period $t$ for the AB customer that starts in position $i$ in the station A queue. Similarly, let
$B_j^{(t)}$, $j = 1, \ldots, N - x^{(t)}$, be the system time in period $t$ for the BA customer that starts in position $j$ in the station $B$ queue. Then, define

$$A^{(t)} := (A_1^{(t)}, \ldots, A_{x^{(t)}}^{(t)}),$$

and

$$B^{(t)} := (B_1^{(t)}, \ldots, B_{N-x^{(t)}}^{(t)}).$$

Customers are capable of reasoning about the system only by means of comparing anecdotes about the system time on each possible route (AB and BA) through the network. Except where specified, we make no specific assumptions regarding how customers choose their routes in the first period. Each new customer in period $t + 1$ randomly polls one AB customer and one BA customer from period $t$ and learns their respective total system times. That is, she takes two independent draws, one from $A^{(t)}$ and one from $B^{(t)}$. If all customers chose the same route in period $t$—that is, either $x^{(t)} = 0$ or $x^{(t)} = N$—then we assume that players in period $t + 1$ follow the route that was unanimously chosen in period $t$. Consequently, “herding” on one route is self-perpetuating, i.e., $x^{(t)} = 0$ implies $x^{(t')}$ = 0 for all $t' \geq t + 1$, and $x^{(t)} = N$ implies $x^{(t')}$ = $N$ for all $t' \geq t + 1$.

If not all customers chose the same route in period $t$, then let $\hat{a}_k^{(t)}$ be the draw from $A^{(t)}$ for customer $k$ in period $t + 1$ (note that $k$ is an arbitrary index and does not relate to the eventual order of service), and similarly $\hat{b}_k^{(t)}$ her draw from $B^{(t)}$. The key to our model of anecdotal reasoning is the following: a customer $k$ in period $t + 1$ will choose route AB if and only if her sample AB system time is no larger than her sample BA system time, that is, if and only if $\hat{a}_k^{(t)}$ $\leq$ $\hat{b}_k^{(t)}$. This form of reasoning is similar to the $S(1)$-reasoning procedure used in Osborne and Rubinstein (1998), Spiegler (2006), and others (see Section 2). Define

$$\hat{a}^{(t)} := (\hat{a}_1^{(t)}, \ldots, \hat{a}_N^{(t)}),$$

and

$$\hat{b}^{(t)} := (\hat{b}_1^{(t)}, \ldots, \hat{b}_N^{(t)}).$$

We can now relate the routing choices in period $t + 1$ to the routing choices in period $t$ through the random draws. Specifically, we can write

$$x^{(t+1)} = \sum_{k=1}^{N} I\{\hat{a}_k^{(t)} \leq \hat{b}_k^{(t)}\}. \quad (1)$$

Letting $p^{(t+1)}$ be the probability that a customer in period $t + 1$ chooses route AB, we have

$$p^{(t+1)} = \Pr[\hat{a}_k^{(t)} \leq \hat{b}_k^{(t)}]. \quad (2)$$
We can infer from equations (1) and (2) that $x^{(t+1)}$ is a binomial random variable with $N$ trials and success probability $p^{(t+1)}$. We wish to express $p^{(t+1)}$ as a function of $x^{(t)}$, and consequently to elucidate the probability distribution of $x^{(t+1)}$. Determining the probability distribution of $x^{(t+1)}$ is computationally burdensome and overly complex analytically. This challenge motivates the use of a fluid model, which eliminates the uncertainty in $x^{(t+1)}$ and smooths out the discreteness. Such fluid models are commonly used to study anecdotal reasoning (see, e.g., Spiegler 2006).

4. A Fluid Model of Open Routing

Hereafter we suppose that customers move through the system as continuous fluid, still making strategic decisions but such that each customer's routing choice has a negligible impact on other customers' system times. Later, in Section 6, we will conduct numerical studies to ascertain how well the insights from our fluid model carry over into the discrete setting.

In the fluid model, a certain volume (normalized to 1) of fluid must be processed, and the entire volume must be processed at both stations. Station $A$ processes fluid at rate $\mu_A$; that is, in an elapsed time of length $\ell$, station $A$ is capable of processing a volume $S_A(\ell)$ of fluid, where

$$S_A(\ell) = \mu_A \ell.$$

Correspondingly, to process a volume $v$ of fluid at station $A$ requires a length of time equal to

$$T_A(v) = \frac{v}{\mu_A}. \quad (3)$$

Station $B$ processes fluid at rate $\mu_B > \mu_A$. So, we similarly have that in an elapsed time of length $\ell$, station $B$ can process a volume of fluid expressed by

$$S_B(\ell) = \mu_B \ell,$$

and processes a volume $v$ of fluid in $T_B(v)$ units of time, where

$$T_B(v) = \frac{v}{\mu_B}. \quad (4)$$

These expressions can be thought of as approximating the limiting case for a discrete system as the number of customers $N$ grows large, if we let the service rates grow proportionally with the number of customers in the system.

Suppose that a fraction $\alpha$ of customers chooses route $AB$. Denote by $Q_A(\alpha; \ell)$ the amount of fluid waiting in the queue (or “buffer”) at station $A$ given an $AB$ fraction of $\alpha$, an elapsed time $\ell$ after the system begins operating. Define $Q_B(\alpha; \ell)$ similarly for station $B$. Thus, we have $Q_A(\alpha; 0) = \alpha,$
and $Q_B(\alpha; 0) = 1 - \alpha$. Letting $[x]^+ = \max\{x, 0\}$, the amount of fluid $Q_B(\alpha; \ell)$ in station $B$’s buffer after the system has been operating for an elapsed time $\ell$ is equal to

$$Q_B(\alpha; \ell) := \left[ Q_B(\alpha; 0) + \min\{Q_A(\alpha; 0), S_A(\ell)\} - S_B(\ell) \right]^+$$

$$= \left[ 1 - \alpha + \min\{\alpha, \mu_A \ell\} - \mu_B \ell \right]^+. \quad (5)$$

This relation takes the positive part of a simple balance equation: after a time $\ell$, the amount in the buffer at station $B$ is equal to the initial quantity, plus the fluid that has arrived from station $A$, minus the fluid that has been processed. Because fluid arrives to station $B$ at a slower rate than it is processed, the quantity in the buffer will be strictly decreasing in $\ell$ until it hits zero, where it remains. We can similarly express the amount of fluid $Q_A(\alpha; \ell)$ in the station $A$ buffer after a time $\ell$ has elapsed by

$$Q_A(\alpha; \ell) := \left[ Q_A(\alpha; 0) + \min\{Q_B(\alpha; 0), S_B(\ell)\} - S_A(\ell) \right]^+$$

$$= \left[ \alpha + \min\{1 - \alpha, \mu_B \ell\} - \mu_A \ell \right]^+. \quad (6)$$

Note that the buffer at station $A$ will first increase with time because fluid arrives to station $A$ faster than it is processed. This increase will continue until the entire volume $1 - \alpha$ of $BA$ customers has departed station $B$ to join the queue at station $A$, after which the station $A$ buffer will shrink at rate $\mu_A$ until it empties.

Let $y_A$ be a possible starting position in the buffer at station $A$, where $0 \leq y_A \leq \alpha$, and let $y_B$ be a possible starting position in the queue at station $B$, where $0 \leq y_B \leq 1 - \alpha$. We denote by $A(\alpha; y_A)$ the total system time for the infinitesimal customer starting in position $y_A$ in the station $A$ queue, given a total $AB$ fraction $\alpha$. Recall that service at station $B$ is strictly faster than at station $A$, i.e., we have $\mu_A < \mu_B$.

**LEMMA 1 (System Time for AB Customers).** The function $A(\alpha; y_A)$ can be expressed as

$$A(\alpha; y_A) = \begin{cases} 
\frac{y_A + 1 - \alpha}{\mu_B} & \text{if } y_A \leq \mu_A \left( \frac{1 - \alpha}{\mu_B - \mu_A} \right), \\
\frac{y_A}{\mu_A} & \text{otherwise.}
\end{cases} \quad (7)$$

The piecewise nature of $A(\alpha; y_A)$ results from taking the positive part of the expression in equation (5) to calculate $Q_B(\alpha; T_A(y_A))$. If $Q_B(\alpha; T_A(y_A))$ is equal to zero, then the fluid at position $y_A$ in the station $A$ buffer will not be delayed in the station $B$ buffer, so the system time is equal to $T_A(y_A)$; this case corresponds to the second piece of equation (7). Otherwise, the fluid will be delayed in the station $B$ buffer, so the total system time will be equal to the time that station $B$
takes to process both its initial buffer of $1 - \alpha$ and a volume of $y_A$ from station $A$’s buffer, hence the first piece of equation (7).

We similarly denote by $B(\alpha; y_B)$ the total system time for the infinitesimal customer starting in position $y_B$ in the station $B$ queue, given a total $AB$ fraction $\alpha$.

**Lemma 2 (System Time for BA Customers).** We can express the function $B(\alpha; y_B)$ by

$$B(\alpha; y_B) = \frac{\alpha + y_B}{\mu_A}. \quad (8)$$

Similar to the discrete customer setting, we suppose that the fluid game is played repeatedly, by new customers in every period. Let $\alpha^{(t)}$ be the fraction of customers who choose route $AB$ in period $t$. The system times of $AB$ customers are represented by an interval $[A^{(t)}, A^{(t)}]$. Similarly, the system times of $BA$ customers are represented by an interval $[B^{(t)}, B^{(t)}]$. We do not impose any restrictions on how customers in the first period choose their routes. Each customer in periods $t \geq 2$ learns the system time of a random $AB$ customer from the previous round: for a customer in period $t + 1$, this system time is a uniform draw from the interval $[A^{(t)}, A^{(t)}]$. Similarly, each customer also learns the system time of a random $BA$ customer from the previous round, whose system time is uniformly distributed on the interval $[B^{(t)}, B^{(t)}]$. We denote by $\hat{a}^{(t)}$ a random draw from the $AB$ interval and $\hat{b}^{(t)}$ a random draw from the $BA$ interval, and these draws are assumed to be independent of each other. Similar to the discrete setting, a customer in period $t + 1$ chooses route $AB$ if and only if $\hat{a}^{(t)} \leq \hat{b}^{(t)}$ (note that because customers are infinitesimal and the draws for all customers are i.i.d., we drop the customer index hereafter).

4.1. Evolution of Customer Behavior Under Anecdotal Reasoning

We seek an equilibrium of this game. Specifically, our notion of equilibrium requires that $\phi(\alpha) = \alpha$, where $\phi(\cdot)$ is the function describing the evolution from $\alpha^{(t)}$ to $\alpha^{(t+1)}$ under anecdotal reasoning. We say that a fraction $\alpha^*$ is an **interior** equilibrium if it is an equilibrium and $0 < \alpha^* < 1$.

It is important to note that our notion of anecdotal reasoning requires a sample system time from both routes. In the event that all customers choose the same route in period $t$—so $\alpha^{(t)} \in \{0, 1\}$—as described in Section 3, we adopt the convention that customers follow the same route in period $t + 1$ (and therefore also in all future periods) that was taken by all customers in period $t$. In other words,
“herding” at one station is self-perpetuating. The values $\alpha = 0$ and $\alpha = 1$ are therefore equilibria by definition. We now develop recursions relating $\alpha^{(t+1)}$ to $\alpha^{(t)}$ for the case with $0 < \alpha^{(t)} < 1$.

**Proposition 1 (Evolution of $\alpha$).** For $t \geq 1$, the sequence $\{\alpha^{(t)}\}$ obeys the recursion

$$
\alpha^{(t+1)} = \phi(\alpha^{(t)}) = \begin{cases}
0 & \text{if } \alpha^{(t)} = 0, \\
1 - \frac{\mu_A}{\mu_B} + \frac{\alpha^{(t)}}{1 - \alpha^{(t)}} \left(1 - \frac{\mu_A}{2\mu_B}\right) & \text{if } 0 < \alpha^{(t)} < \frac{\mu_A}{\mu_A + \mu_B}, \\
1 + \frac{1}{1 - \alpha^{(t)}} - \frac{\mu_A^2 + (\alpha^{(t)})^2\mu_B^2}{2\alpha^{(t)}(1 - \alpha^{(t)})\mu_B} & \text{if } \frac{\mu_A}{\mu_A + \mu_B} \leq \alpha^{(t)} \leq \frac{\mu_A}{\mu_B}, \\
1 & \text{if } \frac{\mu_A}{\mu_B} < \alpha^{(t)} \leq 1.
\end{cases}
$$

(9)

We refer to $\alpha^{(t+1)}$ and $\phi(\alpha^{(t)})$ interchangeably, adopting whichever form is clearer in the present context. Note that, although it is piecewise, $\alpha^{(t+1)}$ is a continuous function of $\alpha^{(t)}$ everywhere except at $\alpha^{(t)} = 0$. Also observe that a value $\alpha$ is an equilibrium if and only if $\phi(\alpha) = \alpha$, i.e., the equilibria occur at fixed points of the function $\phi$.

We now give an intuitive explanation for the piecewise nature of the recursion (9). If $\alpha^{(t)} = 0$ or 1, then the fraction remains the same in the next period. Then, the middle two pieces of the function are delineated based on the system times of the best-off $AB$ customer (i.e., $\hat{y}^{(t)}_A = 0$) and the best-off $BA$ customer ($y^{(t)}_B = 0$). If $0 < \alpha^{(t)} < \mu_A/(\mu_A + \mu_B)$, then the fraction on route $AB$ is small enough that starting at the front of the station $B$ queue yields a shorter system time than starting at the front of the station $A$ queue. In this case, no matter what sample is drawn for $\hat{y}^{(t)}_A$, we have $\Pr[\hat{a}^{(t)} \leq \hat{b}^{(t)} | \hat{y}^{(t)}_A] < 1$ because if the draw for $\hat{y}^{(t)}_B$ is from close enough to the front of the queue, then we will have $\hat{a}^{(t)} < \hat{b}^{(t)}$. By contrast, if $\mu_A/(\mu_A + \mu_B) \leq \alpha^{(t)} \leq \mu_A/\mu_B$, then the best-off $AB$ customer is better off than the best-off $BA$ customer. Thus, there is an interval of $\hat{y}^{(t)}_A$ such that $\Pr[\hat{a}^{(t)} < \hat{b}^{(t)} | \hat{y}^{(t)}_A] = 1$. The existence or nonexistence of this interval leads to the different functional forms of the middle two expressions in equation (9). Finally, if $\mu_A/\mu_B < \alpha^{(t)} < 1$, then even the worst-off $AB$ customer in period-$t$ faces a shorter system time than the very best-off $BA$ customer; thus, we will have $\hat{a}^{(t)} \leq \hat{b}^{(t)}$ for all customers in period $t + 1$, and $\alpha^{(t+1)} = 1$.

Before we further discuss the analytical features of the function $\phi$, we first present some numerical examples to illustrate properties of the function and the possible equilibria. See Figure 2 for two examples of the shape of $\alpha^{(t+1)}$ as a function of $\alpha^{(t)}$, for two different values of $\mu_A/\mu_B$. As will be verified analytically later, these examples are representative in that the function (i) is increasing and (ii) has a convex region followed by a concave region. Equilibria occur where the $\alpha^{(t+1)}$ curve intersects with the line $f(\alpha^{(t)}) = \alpha^{(t)}$, which is also plotted as a visual aid. First, in both plotted cases we have two equilibria, both of which occur in the initial convex interval of the piecewise
recursion for $\alpha^{(t+1)}$. Second, we observe that even for a relatively small change in the service rate ratio (from .95 to .98), there is a substantial change in the equilibrium fractions. Specifically, as the service rate ratio approaches 1, the equilibria spread out toward the extremes of 0 and .5.

Also important is the shape of $\alpha^{(t+1)}$ on either side of the two equilibria. Around the smaller equilibrium (at $\alpha \approx .04$ in the left panel and $\alpha \approx .13$ in the right panel), the curve is crossing from above the line $f(\alpha^{(t)}) = \alpha^{(t)}$ to below it. If $\alpha^{(t+1)}$ is above the line $f(\alpha^{(t)}) = \alpha^{(t)}$, then we have $\alpha^{(t+1)} > \alpha^{(t)}$; this inequality is reversed if $\alpha^{(t+1)}$ is below this line. So, for values of $\alpha^{(t)}$ just below the smaller equilibrium, in the next period $\alpha$ will increase towards the equilibrium value. Similarly, for values just above the smaller equilibrium value, in the next period $\alpha$ will decrease towards the equilibrium. In short, this equilibrium is “stable” in the sense that small perturbations will self-correct back to the equilibrium $\alpha$. The same cannot be said for the larger equilibrium (at $\alpha \approx .47$ in the left panel and $\alpha \approx .40$ in the right panel): values just above or just below it will be pushed away from this equilibrium in the next period. These observations—which will be formalized below—suggest the notion that the smaller of the two interior equilibria may be considered the focal equilibrium.

Now, using the recursion derived in Proposition 1, we find the equilibria by solving a fixed point equation for $\alpha$. Let

$$\eta := 4\sqrt{3} - 6 \approx .9282,$$  \hspace{1cm} (10)

and

$$\nu := \frac{\sqrt{\mu_A^2 + 12\mu_A\mu_B - 12\mu_B^2}}{4\mu_B}.$$  \hspace{1cm} (11)
Proposition 2 (Existence and Multiplicity of Interior Equilibria). Define $\alpha^+_{-}$ and $\alpha^+_{+}$ by

$$\alpha^+_{-} := \frac{1}{2} - \frac{\mu_A}{4\mu_B} - \nu \quad \text{and} \quad \alpha^+_{+} := \frac{1}{2} - \frac{\mu_A}{4\mu_B} + \nu. \quad (12)$$

The number of interior equilibria is either zero, one, or two, and any interior equilibrium $\alpha^*$ must satisfy $0 < \alpha^* < \mu_A/(\mu_A + \mu_B)$. Furthermore, letting $\emptyset$ denote the empty set, the set $\mathcal{E}$ of interior equilibria is given by

$$\mathcal{E} = \begin{cases} 
\emptyset & \text{if } 0 < \mu_A/\mu_B < \eta, \\
\{2 - \sqrt{3}\} & \text{if } \mu_A/\mu_B = \eta, \\
\{\alpha^+_{-}, \alpha^+_{+}\} & \text{if } \eta < \mu_A/\mu_B < 1.
\end{cases}$$

So, depending on the ratio of the service rates, the number of interior equilibria is either zero, one, or two. For ratios of the service rates that are close to one (Figure 2), there are two interior equilibria, both of which occur below $\alpha^{(t)} = .5$. As the service rates become farther apart (Figure 3), the convex portion of the curve increases away from the line $f(\alpha^{(t)}) = \alpha^{(t)}$ (recall that equilibria occur where $\alpha^{(t+1)}$ intersects with this line). When $\mu_A/\mu_B = \eta$, there is only one intersection point, and for $\mu_A/\mu_B < \eta$, there is no intersection and hence, no interior equilibrium. Moreover, there is an abrupt transition at $\mu_A/\mu_B = \eta$. For $\mu_A/\mu_B < \eta$, the only equilibria are the herding equilibria at the extremes of $\alpha^{(t)} = 0$ or 1. But as soon as $\mu_A/\mu_B \geq \eta$, immediately there are one or more interior equilibria which are nowhere near the herding equilibria. For instance, at $\mu_A/\mu_B = \eta$, there is one interior equilibrium at $\alpha^{(t)} = 2 - \sqrt{3} \approx .27$ (left panel of Figure 3). Then, as the ratio approaches one, the interior equilibria spread out towards $\alpha^{(t)} = 0$ and 1/2 (left panel of Figure 2).
Thus, we find that when the service rates are relatively close together, equilibria emerge under customer anecdotal reasoning which depart significantly from the predicted behavior of fully rational customers. It is important to assess (i) whether these interior equilibria are robust, i.e., if customers can reasonably be expected to arrive at them, and (ii) how well the system performs at these interior equilibria, relative to the fully rational and socially optimal cases. These issues will be addressed in Section 4.2 and Section 5, respectively.

4.2. Convergence to Interior Equilibrium

We have proven the existence of interior equilibria for service rates that are not too far apart. However, the question naturally arises of the stability (or instability) of these equilibria. We address this question by showing that for a range of initial states, the path of play in the fluid model converges to $\alpha^*_-$, the smaller of the two interior equilibrium fractions. Specifically, we suppose that the game is played repeatedly and that customers in each round reason anecdotally, so that play evolves from one period to the next according to equation (9). For rounds $t = 1, 2, \ldots$, we denote this sequence by $\{\alpha^{(t)}\}$. We have the following convergence result.

**Proposition 3 (Convergence to Interior Equilibrium $\alpha^*_-$).** If $\mu_A/\mu_B \geq \eta$ and $0 < \alpha^{(1)} < \alpha^*_+$, then $\{\alpha^{(t)}\}$ converges to $\alpha^*_-$, that is, we have $\lim_{t \to \infty} \alpha^{(t)} = \alpha^*_-$.

Thus, although there can be two interior equilibria, the focal equilibrium is the smaller $\alpha^*_-$. This point was argued graphically earlier based on Figure 2, and Proposition 3 confirms the intuition. The result further supports the notion that the interior equilibria under anecdotal reasoning are more than mere computational artifacts. Rather, for a wide range of initial conditions, including starting points both above and below $\alpha^*_-$, customers who reason anecdotally will converge to the interior equilibrium $\alpha^*_-$. This result stands in marked contrast to the behavior of fully rational customers, who will converge to herding under a range of initial conditions (Arlotto et al. 2019, Proposition 3). Moreover, not only does this represent a regime in which the routing choices of customers differ under anecdotal reasoning from the fully rational case, but also the choices under anecdotal reasoning can make them substantially worse off individually and as a group, as we will see in the next section.

By contrast, if the initial fraction of $AB$ customers is above the larger interior equilibrium fraction $\alpha^*_+$, or if the service rates admit no interior equilibria, then play converges to herding on route $AB$.

**Proposition 4 (Convergence to Herding).** If either (i) $\mu_A/\mu_B \geq \eta$ and $\alpha^{(1)} > \alpha^*_+$, or (ii) $\mu_A/\mu_B < \eta$ and $0 < \alpha^{(1)} \leq 1$, then $\{\alpha^{(t)}\}$ converges to 1, that is, we have $\lim_{t \to \infty} \alpha^{(t)} = 1$. 

The reason for part (i) of this result is related to the instability of the larger equilibrium $\alpha^*_+$: starting at any $\alpha^{(1)} > \alpha^*_+$, the sequence $\alpha^{(i)}$ is increasing in $t$, and it eventually converges to 1. Similarly for part (ii), if $\mu_A/\mu_B < \eta$, then there are no interior equilibria, and the sequence $\{\alpha^{(i)}\}$ is increasing in $t$ for any $0 < \alpha^{(1)} \leq 1$ (see the right panel of Figure 3). Propositions 3 and 4 together suggest an unusual dichotomy. Namely, under customer anecdotal reasoning, there exist two regimes: one in which customers converge to an outcome completely unrelated to that chosen by fully rational customers, and another in which anecdotally reasoning customers are indistinguishable from fully rational ones at equilibrium.

5. Cumulative System Time

Having now a sound understanding of customer equilibrium routing decisions under anecdotal reasoning, we next compare the performance of different systems.

Service providers serving delay-sensitive customers may want to reduce the delay faced by customers, as those who endure overlong waits may not return to purchase again in the future. Similarly, customer willingness to pay may be reduced because of the cost associated with waiting and/or a lower perceived quality.

To measure how well the system performs—and by extension, how satisfied customers are likely to be—we use the cumulative system time. This quantity measures the cumulative delay experienced by customers by the integral of the total system time experienced by customers in each position in the queues. The cumulative system time is derived in the following proposition.

**Proposition 5 (Cumulative System Time).** Let

$$\xi_\alpha := \min \left\{ \alpha, \mu_A \left( \frac{1-\alpha}{\mu_B - \mu_A} \right) \right\}. \quad (13)$$

The cumulative system time is given by

$$D(\alpha, \mu_A, \mu_B) = \begin{cases} \frac{\alpha(1-\alpha)}{\mu_A} + \frac{(1-\alpha)^2}{2\mu_A} + \frac{\alpha(1-\alpha)}{\mu_B} + \frac{\alpha^2}{2\mu_B} & \text{if } \alpha \leq \frac{\mu_A}{\mu_B}, \\ \frac{\alpha(1-\alpha)}{\mu_A} + \frac{(1-\alpha)^2}{2\mu_A} + \frac{\xi_\alpha(1-\alpha)}{\mu_B} + \frac{\alpha^2}{2\mu_A} + \frac{\xi_\alpha^2}{2\mu_B} \left( \frac{1}{2\mu_B} - \frac{1}{2\mu_A} \right) & \text{otherwise}. \end{cases} \quad (14)$$

We can deduce from equation (14) that $D(0, \mu_A, \mu_B) = D(1, \mu_A, \mu_B)$, i.e., herding on route $AB$ yields the same cumulative system time as herding on route $BA$. The cumulative system time $D(\cdot)$ is plotted as a function of $\alpha$ for fixed values of $\mu_A$ and $\mu_B$ in Figure 4. We observe in these plots that, for a fixed ratio of the service rates, the smallest possible cumulative system time tends to occur when all customers follow the same route (i.e., $\alpha = 0$ or $\alpha = 1$).
Next, to understand how the cumulative system time varies with the service rates, we fix the sum of the service rates at 1. That is, we study the cumulative system time as a function of $x$, where $\mu_A = x$ and $\mu_B = 1 - x$. We will compare the system performance in equilibrium between the fully rational case and the anecdotal reasoning case. It is convenient to relate $x$ to the service rate ratio $r = \mu_A / \mu_B$. We have

$$r = \frac{x}{1 - x},$$

and therefore

$$x = \frac{r}{1 + r}.$$ 

As $x$ ranges from 0 to $1/2$, the ratio $r$ ranges from 0 to 1.

See Figure 5, which depicts the cumulative system time at each of the different equilibria: the region with $r < 1/2$ is not plotted because the cumulative system time under herding grows without bound as the ratio approaches zero from the left (there are no interior equilibria in this region). With these provisos, the figure covers all possible cases with $\mu_A = x < 1/2$ and $\mu_B = 1 - x$. For the anecdotal reasoning case, if $\mu_A / \mu_B < \eta$, then there is no interior equilibrium, and the equilibrium is for customers to herd, just like the fully rational case. Thus, in this regime the system performs the same regardless of the form of reasoning used by customers. By contrast, in the intermediate regime where $\eta < \mu_A / \mu_B < 1$, at the interior anecdotal reasoning equilibrium $\alpha^*$, the system performs strictly worse than under herding. Define

$$G(\alpha, r) = \frac{D(\alpha, \frac{r}{1+r}, \frac{1}{1+r}) - D(0, \frac{r}{1+r}, \frac{1}{1+r})}{D(0, \frac{r}{1+r}, \frac{1}{1+r})},$$
the percentage excess cumulative system time at the equilibrium $\alpha^*$, as compared to herding (note that as mentioned, the cumulative system time is the same for herding on route $AB$ as for herding on route $BA$). As $\mu_A/\mu_B$ moves above $\eta \approx .9282$, we move from a region with no interior equilibria into a region with two, and we immediately see the performance suffer. As seen in panel (b), the cumulative system time at $\alpha^*$ can be as much as 35% worse than that under herding; that is, $G(\alpha^*, r)$ can be as large as .35. As the service rate ratio approaches 1, however, the interior equilibrium $\alpha^*$ approaches 0 (i.e., herding on route $BA$), and the degradation in system performance diminishes. Thus we observe an unusual phenomenon, namely that the performance loss when customers use anecdotal reasoning instead of being fully rational is non-monotonic in the ratio of the service rates. When the service rates are far apart, the equilibria under anecdotal reasoning align perfectly with the fully rational case, and there is no performance loss. When the rates approach each other, the equilibrium $\alpha^*$ approaches herding on route $BA$, and the performance loss vanishes (recall that the larger interior equilibrium $\alpha^*_+$ is highly unstable and unlikely to be implemented, much less maintained). It is in the intermediate region when the service rates are close, but not too close, that the performance of the system suffers the most.

These observations lead to an unusual prescription if service providers are able to choose the service rates. Customer behavior differs significantly under anecdotal reasoning from the fully rational case, and for some service rates the performance of the system is far worse. However, in both cases, as can be seen from Figure 5, it is best to set the service rates as close to each other
as possible. The reason is that as the gap between the service rates narrows, customer routing decisions under anecdotal reasoning begin to approximate herding, and when customers herd it is best to balance the service rates at the two stations.

6. Systems with Atomic Customers

We now conduct a simulation study of systems with atomic customers, with the goal of ascertaining how closely the behavior of these systems resembles the fluid model. For a range of initial fractions $\alpha^{(1)}$ and service rate ratios $\mu_A/\mu_B$, we simulated systems with $N = 500$ customers for 100 rounds each, replicating each system 10 times. We make two key observations regarding the output of the study.

First, we observe that, even for systems with only 500 customers, our analytical results for the fluid system are relatively good predictors of the evolution of customer routing decisions. Table 1 reports sample statistics fraction of $AB$ customers in rounds 75-100, across replications for each parameter combination. For reference, the value of the smaller fluid equilibrium $\alpha^-_*$ is given in the first row of the table. We observe that when the initial fraction $\alpha^{(1)}$ is reasonably small—more precisely, when it is smaller than the larger interior equilibrium $\alpha^+_*$—play tends to hover very closely around the fluid equilibrium. The mean and median are extremely close to $\alpha^-_*$, and the first and third quartile are roughly .01 above or below $\alpha^-_*$, respectively. Figure 6 shows two typical sample paths of the discrete system. The initial $AB$ fraction is between the interior equilibria. In both sample paths, play quickly approaches the fluid equilibrium and continues to hover around it, with the amplitude of the variations decreasing over time. We observe similar outcomes commonly, particularly when $\alpha^{(1)}$ is significantly smaller than $\alpha^+_*$.

While the first observation pertained to the similarities between the discrete and fluid systems, the second highlights their differences. Unlike the fluid system (whose evolution is depicted as a dashed line), in the discrete system it is possible to start with $\alpha^{(1)}$ between the interior equilibria and still converge to herding at $\alpha = 1$. This situation is depicted in Figure 7, which shows two sample paths with the same initial fraction $\alpha^{(1)}$ as in Figure 6, but in which customers converge to herding on route $AB$. This departure from the fluid equilibrium can be ascribed to the randomness in the sampling: if a disproportionate number of customers receive samples suggesting that route $AB$ provides a shorter system time, then a cascading effect can occur where $\alpha$ “escapes” from the region between the equilibria and then increases inexorably to 1 over time. In panel (a), we see this escape happen almost immediately. Panel (b) even more powerfully exhibits the role of the randomness: customers first move above the higher equilibrium $\alpha^+_*$, suggesting that play will
Table 1  Statistics for fraction of $AB$ customers in rounds 75-100 ($N = 500$, 100 total rounds, 10 replications)

<table>
<thead>
<tr>
<th>$\alpha^*$</th>
<th>$\mu_A/\mu_B$</th>
<th>$\mu_A/\mu_B$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>1.069 0.995 0.042 0.004</td>
<td>0.164 0.094 0.043 0.010</td>
</tr>
<tr>
<td>0.15</td>
<td>1.068 0.997 0.043 0.006</td>
<td>0.166 0.094 0.042 0.004</td>
</tr>
<tr>
<td>0.25</td>
<td>0.918 0.092 0.041 0.007</td>
<td>0.178 0.092 0.042 0.004</td>
</tr>
<tr>
<td>0.35</td>
<td>1.056 0.01</td>
<td>0.144 0.044 0.004</td>
</tr>
</tbody>
</table>

(a) Mean  

<table>
<thead>
<tr>
<th>$\alpha^*$</th>
<th>$\mu_A/\mu_B$</th>
<th>$\mu_A/\mu_B$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>1.050 0.082 0.034 0</td>
<td>0.150 0.082 0.034 0</td>
</tr>
<tr>
<td>0.15</td>
<td>1.054 0.084 0.034 0</td>
<td>0.154 0.084 0.034 0</td>
</tr>
<tr>
<td>0.25</td>
<td>1.050 0.082 0.034 0</td>
<td>0.194 0.106 0.0465 0</td>
</tr>
<tr>
<td>0.35</td>
<td>1.056 0.036 0</td>
<td>1.014 0.104 0.05 0.008</td>
</tr>
<tr>
<td>0.45</td>
<td>1.056 0.01</td>
<td>1.045 0.056 0.01</td>
</tr>
</tbody>
</table>

(b) Median  

<table>
<thead>
<tr>
<th>$\alpha^*$</th>
<th>$\mu_A/\mu_B$</th>
<th>$\mu_A/\mu_B$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>1.088 0.106 0.048 0.008</td>
<td></td>
</tr>
<tr>
<td>0.15</td>
<td>0.182 0.104 0.05 0.008</td>
<td></td>
</tr>
<tr>
<td>0.25</td>
<td>0.194 0.106 0.0465 0</td>
<td></td>
</tr>
<tr>
<td>0.35</td>
<td>1.045 0.056 0.01</td>
<td></td>
</tr>
</tbody>
</table>

(c) First Quartile  

<table>
<thead>
<tr>
<th>$\alpha^*$</th>
<th>$\mu_A/\mu_B$</th>
<th>$\mu_A/\mu_B$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>0.164 0.094 0.043 0.010</td>
<td></td>
</tr>
<tr>
<td>0.15</td>
<td>0.166 0.094 0.042 0.004</td>
<td></td>
</tr>
<tr>
<td>0.25</td>
<td>0.172 0.096 0.042 0.004</td>
<td></td>
</tr>
<tr>
<td>0.35</td>
<td>0.182 0.092 0.042 0.004</td>
<td></td>
</tr>
<tr>
<td>0.45</td>
<td>0.178 0.092 0.042 0.004</td>
<td></td>
</tr>
</tbody>
</table>

(d) Third Quartile

---

Figure 6  Simulated evolution of discrete system ($N = 500$, $\mu_A = .94$, $\mu_B = 1$).

likely converge to herding on route $AB$ as in panel (a). However, the evolution reverses course to move back between the equilibria, only to escape again and eventually converge to herding. We see this observation borne out in Table 1; when $\alpha^{(1)}$ is large, play more often converges to herding on route $AB$, leading to values of 1 in the bottom rows of Table 1. Thus, while the fluid model is a good predictor of the “average” behavior of the discrete system, the randomness can still result in unusual outcomes. Finally, we note that if the initial fraction $\alpha^{(1)} \geq \alpha^*_*$, then the fluid system will converge to herding on route $AB$, and the discrete system will usually do the same.
7. The Case with Multiple Samples

So far we have focused on the $S(1)$ variant of anecdotal reasoning, in which customers receive only a single system-time sample for each route. However, it is plausible that customers might receive multiple samples for each route and aggregate them in some way to form quasi-beliefs about the expected system time for each route.

Specifically, suppose that customers receive $n \geq 2$ samples on each route and take the sample average of these, yielding estimates $\hat{a}_n^{(t)}$ and $\hat{b}_n^{(t)}$ of the expected system time on route $AB$ and $BA$, respectively. A customer in period $t + 1$ will choose route $AB$ if and only if $\hat{a}_n^{(t)} \leq \hat{b}_n^{(t)}$. Denote by $\alpha_n^{(t)}$ the fraction of customers choosing route $AB$ in period $t$, given that customers receive $n$ samples. We can compute numerically the exact value of $\alpha_n^{(t)}$ for small values of $n$; examples are plotted in Figure 8 and Figure 9a. For larger values of $n$, while it is computationally prohibitive to calculate $\alpha_n^{(t)}$ exactly, we can instead apply the central limit theorem to approximate it. Customers will be taking the sample average of several independent and identically distributed random variables as their estimates for routes $AB$ and $BA$, and then comparing these sample averages. The difference between the sample averages is the difference between two independent, approximately normal random variables, which itself will be approximately normal. We can thus compute an approximate value for $\alpha_n^{(t)}$ as the value of the distribution function at zero for the normal approximation of the difference between the sample-average $AB$ system time and the sample-average $BA$ system time.

Figure 8 depicts the evolution of $\alpha_n^{(t)}$ for $n = 2$ and $n = 3$, for the normal approximation as well as the exact evolution. We can see that, even with multiple samples, there are interior (that is, non-herding) equilibria. The equilibria for $n = 2$ and $n = 3$ differ from the single-sample interior equilibria and from each other, but the shapes of $\alpha_n^{(t)}$ are both similar to the one-sample case, i.e.,
increasing with a convex interval followed by a concave interval. We have the ratio $\mu_A/\mu_B = .84$ in Figures 8 and 9, which is substantially smaller than the one-sample boundary $\eta \approx .9282$, and still, there are two interior equilibria with $n = 2$, $n = 3$, and $n = 5$. So, with $n = 2$, there appears to be a wider interval of $\mu_A/\mu_B$ for which an interior equilibrium exists than with $n = 1$. Moreover, the existence of interior equilibria for $n = 3$ and $n = 5$, together with the shifting of the $\alpha^{(t)}_n$-curve to the right for larger $n$, suggests that the interval with interior equilibria continues to grow as $n$ grows.

As mentioned, it becomes computationally prohibitive to calculate $\alpha^{(t)}_n$ for more than a few samples, so we instead resort to a normal approximation which becomes more accurate as $n$ increases. For $n = 2$ and $n = 3$, the normal approximation is not all that accurate (see Figure 8), but as $n$ grows, the two curves converge. Moreover, customer decisions do not change much once the number of samples increases beyond 5. We can see in Figure 9 that the curve for the normal approximation is already almost as steep with $n = 5$ as it will be with $n = 100$, and the curves for both $n = 5$ and $n = 100$ begin their steep ascent at around the same value of $\alpha^{(t)}_n$.

As the number of samples grows, the functional form of $\alpha^{(t)}_n$ becomes steeper. This phenomenon implies that the “intermediate region” where customers could possibly converge to an interior equilibrium becomes smaller and smaller. Customers are then pushed ever more inevitably towards the two extremes of $\alpha = 0$ or 1, i.e., herding as in the fully rational case. This limiting behavior is not surprising and is consistent with other findings in the anecdotal reasoning literature, where as
customers take more samples their behavior approaches that under full rationality (see, e.g., Ren et al. 2018).

8. Conclusion

We have studied the evolution and equilibrium dynamics of an open-routing service network with self-interested, boundedly rational customers who are limited in their ability to reason about the system. We fully characterized the evolution of customer routing decisions over time in a fluid model. The explicit expressions for the sequence of routing profiles chosen allowed us to completely determine the set of equilibria, in particular “interior equilibria which involve a split of customers among the two possible routes through the network.

When the ratio of the service rates at the two stations is either relatively small or very close to one, we found that in equilibrium our anecdotally reasoning customers behave similarly to the fully rational case: they herd at one station and take the same route through the network. But between these two regimes, there is a range for the service rate ratio wherein equilibria exist which do not involve herding at one station. Moreover, one particular interior equilibrium is quite robust: for a range of initial conditions, we proved that customers will converge to this equilibrium rather than to herding. At these interior equilibria, the average time that customers spend in the system can balloon to as much as 35% larger than under herding.

In addition, we ran an extensive numerical study to uncover the extent to which our results for the fluid model approximate systems with discrete customers. We found that for systems with even
moderately many customers, the fluid results were a good approximation: play usually converged very close to the predicted interior equilibria under the initial conditions hypothesized by our analytical results. Finally, we extended our study to systems where customers employ anecdotal reasoning with more than one sample. The output of a numerical study showed that interior equilibria continue to exist with two or more samples, with customers behaving more and more like fully rational players as the number of samples increased.

The foregoing discussion might suggest that with customer anecdotal reasoning, a service provider who can choose the service rates might make a different decision from a service provider with fully rational customers—e.g., he might choose to differentiate the service rates so as to avoid the regime in which customers can converge to a non-herding equilibrium. This intuition is misleading, however. The average system time for the equilibrium to which customers converge can indeed be quite bad. At the same time, when the service rates approach one another, the interior equilibrium to which customers converge approaches herding and the optimal system performance. Thus, even though the equilibrium customer routing decisions with anecdotal reasoning can differ considerably from the fully-rational setting, in both cases it is best to set the service rates as close to each other as possible.

Taken together, our results suggest that service providers should attend closely to customer routing decisions and, if possible, gather information on how customers choose their routes. For systems where the provider has less control over the service rates, customers who reason anecdotally can cause themselves to wait much longer than necessary, but these bad outcomes may be avoided by educating customers. If the provider can indeed control the service rates, then whether customers are fully rational or they reason anecdotally, he would be wise to set the service rates as close together as possible to ensure the smallest achievable average waiting time.

Appendix. Proofs of Technical Results

Proof of Lemma 1. The fluid at position $y_A$ in the buffer at station $A$ will depart from station $A$ after $T_A(y_A) = y_A/\mu_A$ units of time by equation (3). When this customer arrives at station $B$, the amount of fluid in the buffer there is $Q_B(\alpha; y_A/\mu_A)$, which we can determine using equation (5). We can therefore write her total system time $A(\alpha; y_A)$ as

$$A(\alpha; y_A) = \frac{y_A}{\mu_A} + T_B\left(Q_B\left(\alpha; \frac{y_A}{\mu_A}\right)\right)$$

$$= \frac{y_A}{\mu_A} + \frac{1}{\mu_B} \left[1 - \alpha + \min\{\alpha, y_A\} - \frac{\mu_B y_A}{\mu_A}\right]^+$$

$$= \frac{y_A}{\mu_A} + \frac{1}{\mu_B} \left[1 - \alpha + y_A - \frac{\mu_B y_A}{\mu_A}\right]^+, \quad (15)$$
when the last substitution follows from the fact that for any \( AB \) customer we must have \( y_A \leq \alpha \). The bracketed term in equation (15) is nonnegative only if \( y_A \leq \mu_A(1-\alpha)/\mu_B - \mu_A \). In this case, substitution yields the first expression in equation (7). Otherwise, i.e., if \( y_A > \mu_A(1-\alpha)/\mu_B - \mu_A \), taking the positive part of the bracketed term gives zero, and we are left with \( y_A/\mu_A \). We conclude equation (7). □

**Proof of Lemma 2.** The proof proceeds along similar lines to the proof of Lemma 1. The fluid at position \( y_B \) in the buffer at station \( B \) will depart from station \( B \) after \( T_B(y_B) = y_B/\mu_B \) units of time by equation (4). When this customer arrives at station \( A \), the amount of fluid in the buffer there is \( Q_A(\alpha; y_B/\mu_B) \), which we can determine using equation (6). The total system time for this customer is then

\[
B(\alpha; y_B) = \frac{y_B}{\mu_B} + T_A \left( Q_A \left( \alpha; \frac{y_B}{\mu_B} \right) \right)
\]

\[
= \frac{y_B}{\mu_B} + \frac{1}{\mu_A} \left[ \alpha + \min \left\{ 1 - \alpha, y_B \right\} - \frac{\mu_A y_B}{\mu_B} \right]^+
\]

\[
= \frac{y_B}{\mu_B} + \frac{1}{\mu_A} \left( \alpha + y_B - \frac{\mu_A y_B}{\mu_B} \right)
\]

\[
= \frac{\alpha + y_B}{\mu_A}.
\]

The equivalence between the second and third equations holds because we must have \( y_B \leq 1 - \alpha \), and because \( \mu_A/\mu_B \leq 1 \) then implies that the bracketed term is nonnegative. □

**Proof of Proposition 1.** We proceed by cases.

**Case 1:** \( \alpha^{(t)} = 0 \). In this case, by assumption we have \( \alpha^{(t+1)} = 0 \).

**Case 2:** \( 0 < \alpha^{(t)} < \mu_A/(\mu_A + \mu_B) \). Straightforward algebra reveals that \( \alpha^{(t)} < \mu_A/(\mu_A + \mu_B) < \mu_A/\mu_B \) implies \( \alpha^{(t)} < \mu_A(1-\alpha^{(t)})/(\mu_B - \mu_A) \). Thus, because \( y_A \leq \alpha^{(t)} \), we can ignore the second piece of the function \( A(\cdot) \) given in equation (7). A customer in period \( t + 1 \) will choose route \( AB \) if and only if her random draws satisfy \( \hat{a}^{(t)} \leq \hat{b}^{(t)} \), that is, if and only if

\[
\frac{\hat{y}_A^{(t)} + 1 - \alpha^{(t)}}{\mu_B} \leq \frac{\alpha^{(t)} + \hat{y}_B^{(t)}}{\mu_A}
\]

\[
\iff \frac{\mu_A}{\mu_B} \left( \hat{y}_A^{(t)} + 1 - \alpha^{(t)} \right) - \alpha^{(t)} \leq \hat{y}_B^{(t)}.
\]

Therefore, the probability that a customer in period \( t + 1 \) chooses route \( AB \) is given by

\[
Pr[\hat{a}^{(t)} \leq \hat{b}^{(t)}] = Pr \left[ \frac{\mu_A}{\mu_B} \left( \hat{y}_A^{(t)} + 1 - \alpha^{(t)} - \frac{\alpha^{(t)} \mu_B}{\mu_A} \right) \leq \hat{y}_B^{(t)} \right].
\]  

(16)

Because the position of the randomly drawn period-\( t \) \( BA \) customer is uniformly distributed within the station \( B \) queue, the random variable \( \hat{y}_B^{(t)} \) has a uniform distribution on the interval \([0, 1 - \alpha^{(t)}] \). Furthermore, \( \alpha^{(t)} < \mu_A/(\mu_A + \mu_B) \) implies

\[
\frac{\mu_A}{\mu_B} (1 - \alpha^{(t)}) - \alpha^{(t)} > 0.
\]

Therefore, conditional on a value of \( \hat{y}_A^{(t)} \), we have

\[
Pr[\hat{a}^{(t)} \leq \hat{b}^{(t)} | \hat{y}_A^{(t)}] = 1 - \frac{\mu_A (\hat{y}_A^{(t)} + 1 - \alpha^{(t)}) - \alpha^{(t)} \mu_B}{\mu_B (1 - \alpha^{(t)})}.
\]  

(17)
Similar to \( \hat{y}_A^{(t)} \), the random variable \( \hat{y}_A^{(t)} \) has a uniform distribution on \([0, \alpha^{(t)}]\). It therefore has density function \( f(x) \), where
\[
f(x) = \begin{cases} \frac{1}{\alpha^{(t)}} & \text{if } 0 \leq x \leq \alpha^{(t)}, \\ 0 & \text{otherwise.} \end{cases}
\]

Integrating over this density function, we compute \( \alpha^{(t+1)} \), the probability that a customer in period \( t+1 \) chooses route \( AB \), by
\[
\alpha^{(t+1)} = \Pr[\hat{a}^{(t)} \leq \hat{b}^{(t)}] \\
= \int_0^{\alpha^{(t)}} \Pr[\hat{a}^{(t)} \leq \hat{b}^{(t)} | \hat{y}_A^{(t)} = x] f(x) dx \\
= \frac{1}{\alpha^{(t)}} \int_0^{\alpha^{(t)}} \left[ 1 - \frac{\mu_A(x + 1 - \alpha^{(t)}) - \alpha^{(t)} \mu_B}{\mu_B(1 - \alpha^{(t)})} \right] dx \\
= 1 - \frac{\mu_A}{\mu_B} + \frac{\alpha^{(t)}}{1 - \alpha^{(t)}} \left( 1 - \frac{\mu_A}{2\mu_B} \right).
\]

**Case 3:** \( \mu_A/(\mu_A + \mu_B) \leq \alpha^{(t)} \leq \mu_A/\mu_B \). The probability that a customer in period \( t+1 \) will choose route \( AB \) is the probability that \( \hat{a}^{(t)} \leq \hat{b}^{(t)} \), and takes the same form as equation (16). If the left-hand side (LHS) of the inequality inside the brackets in equation (16) is negative, then the probability is necessarily equal to 1 because \( \hat{y}_B^{(t)} \) takes only nonnegative values. If the LHS of the inequality is nonnegative, then the probability is equal to the expression given in equation (17). Specifically, conditional on a value of \( \hat{y}_A^{(t)} \) such that \( 0 \leq \hat{y}_A^{(t)} \leq \alpha^{(t)} \), we have
\[
\Pr[\hat{a}^{(t)} \leq \hat{b}^{(t)} | \hat{y}_A^{(t)}] = \begin{cases} 1 & \text{if } \hat{y}_A^{(t)} + 1 - \alpha^{(t)} - \frac{\alpha^{(t)} \mu_B }{\mu_A} \leq 0, \\ 1 - \frac{\mu_A(\hat{y}_A^{(t)} + 1 - \alpha^{(t)}) - \alpha^{(t)} \mu_B}{\mu_B(1 - \alpha^{(t)})} & \text{otherwise.} \end{cases}
\]

The remainder of the proof is exactly analogous to the proof of Case 1, with the end result being the recursion in the third piece of equation (9).

**Case 4:** \( \mu_A/\mu_B < \alpha^{(t)} < 1 \). If \( \mu_A/\mu_B < \alpha^{(t)} < 1 \), then for \( 0 \leq y_A \leq \alpha^{(t)} \), we have
\[
\mathcal{A}(y_A) \leq A(\alpha^{(t)}) = \frac{\alpha^{(t)}}{\mu_A},
\]
and
\[
\mathcal{B}(y_B) = \frac{\alpha^{(t)} + y_B}{\mu_A} \geq \frac{\alpha^{(t)}}{\mu_A} = A(\alpha^{(t)}) \text{ for } 0 \leq y_B \leq 1 - \alpha^{(t)}.
\]
Thus, for any possible draw \( \hat{y}_A^{(t)} \) for route \( AB \) and any possible draw \( \hat{y}_B^{(t)} \) for route \( BA \), we have \( \mathcal{A}(\hat{y}_A^{(t)}) \leq \mathcal{B}(\hat{y}_B^{(t)}) \). Thus, for \( \mu_A/\mu_B < \alpha^{(t)} < 1 \), we have
\[
\alpha^{(t+1)} = \Pr[\hat{a}^{(t)} \leq \hat{b}^{(t)}] = 1.
\]

**Case 5:** \( \alpha^{(t)} = 1 \). In this case, by assumption we have \( \alpha^{(t)} = 1 \).

Combining Cases 1-5, we conclude equation (9).
Proof of Proposition 2. The proof of Proposition 2 proceeds through several lemmas. First, Proposition 1 rules out the possibility of an interior equilibrium in the interval \((\mu_A/\mu_B, 1)\). Next, the following two lemmas establish some useful properties of the recursion which will be used later in the proof.

**Lemma 3 (Initial Convexity of \(\alpha\)).** The fraction \(\alpha^{(t+1)} = \phi(\alpha^{(t)})\) is strictly convex and strictly increasing in \(\alpha^{(t)}\) on the interval \((0, \mu_A/(\mu_A + \mu_B))\).

**Proof of Lemma 3.** Let \(\alpha^{(t+1)} = \phi(\alpha^{(t)})\). Then, for \(0 < \alpha^{(t)} < \mu_A/(\mu_A + \mu_B)\), we have
\[
\alpha^{(t+1)} = \phi(\alpha^{(t)}) = 1 - \frac{\mu_A}{\mu_B} + \frac{\alpha^{(t)}}{1 - \alpha^{(t)}} \left(1 - \frac{\mu_A}{2\mu_B}\right).
\]
Define \(c := 1 - \mu_A/2\mu_B > 0\). Differentiating with respect to \(\alpha^{(t)}\), we have
\[
\phi'(\alpha^{(t)}) = c \left(\frac{1}{1 - \alpha^{(t)}} + \frac{\alpha^{(t)}}{1 - \alpha^{(t)}}\right) > 0,
\]
where the strict positivity follows because \(c > 0\) and by assumption \(0 < \alpha^{(t)} < \mu_A/(\mu_A + \mu_B)\). The function is therefore strictly increasing on the relevant interval. Differentiating a second time, we get
\[
\phi''(\alpha^{(t)}) = c \left(\frac{2\alpha^{(t)}}{(1 - \alpha^{(t)})^3} + \frac{2}{(1 - \alpha^{(t)})^2}\right) > 0,
\]
where strict positivity follows for the same reasons as for the first derivative. Using the second-order condition for convexity (Boyd and Vandenberghe 2004, Section 3.1.4), we conclude that \(\alpha^{(t+1)} = \phi(\alpha^{(t)})\) is strictly convex in \(\alpha^{(t)}\) on the interval \((0, \mu_A/(\mu_A + \mu_B))\). □

**Lemma 4 (Later Concavity of \(\alpha\)).** The fraction \(\alpha^{(t+1)} = \phi(\alpha^{(t)})\) is increasing and strictly concave in \(\alpha^{(t)}\) on the interval \([\mu_A/(\mu_A + \mu_B), \mu_A/\mu_B]\).

**Proof of Lemma 4.** Differentiating twice, we get
\[
\phi''(\alpha^{(t)}) = \frac{(\mu_B - \mu_A)^2}{\mu_A \mu_B (\alpha^{(t)} - 1)^3} - \frac{\mu_A}{\mu_B (\alpha^{(t)})^3}.
\]
(19)
The first term in equation (19) is strictly negative because \(\mu_A < \mu_B\): the numerator is positive because it is a square, and the denominator is strictly negative because \(0 < \alpha^{(t)} < 1\) in the range we are considering. The second term in equation (19) is strictly positive because \(\mu_A, \mu_B, \) and \(\alpha^{(t)}\) are all strictly positive. Thus, equation (19) entails subtracting a positive number from a negative one, and we conclude
\[
\phi''(\alpha^{(t)}) = \frac{(\mu_B - \mu_A)^2}{\mu_A \mu_B (\alpha^{(t)} - 1)^3} - \frac{\mu_A}{\mu_B (\alpha^{(t)})^3} < 0.
\]
Therefore, the function \(\phi(\alpha^{(t)})\) is strictly concave for \(\mu_A/(\mu_A + \mu_B) \leq \alpha^{(t)} \leq \mu_A/\mu_B\) by the second-order condition for concavity. The function \(\phi(\alpha^{(t)})\) is only semi-differentiable at \(\alpha^{(t)} = \mu_A/(\mu_A + \mu_B)\); that is, it has left and right derivatives at this point—both first and second—but they are not equal to each other. We use the right derivative at this point to establish concavity on the closed interval \([\mu_A/(\mu_A + \mu_B), \mu_A/\mu_B]\).

Now, the fact that the second derivative \(\phi''(\alpha^{(t)}) < 0\) implies that the first derivative is strictly decreasing, i.e., for \(\mu_A/(\mu_A + \mu_B) \leq \alpha^{(t)} < \hat{\alpha}^{(t)} \leq \mu_A/\mu_B\), we have
\[
\phi'(\alpha^{(t)}) > \phi'(\hat{\alpha}^{(t)}).
\]
Evaluating the first derivative gives

\[ \phi'(\alpha(t)) = \frac{-(\mu_B - \mu_A)^2}{2\mu_A\mu_B(1-\alpha(t))^2} + \frac{\mu_A}{2\mu_B(\alpha(t))^2}. \]  

(20)

Substituting \( \alpha(t) = \mu_A/\mu_B \) gives

\[ \phi'(\frac{\mu_A}{\mu_B}) = 0, \]

which, together with the fact that the first derivative is strictly decreasing on this interval, implies that

\[ \phi'(\alpha(t)) \geq 0 \quad \text{for} \quad \frac{\mu_A}{\mu_A + \mu_B} \leq \alpha(t) \leq \frac{\mu_A}{\mu_B}. \]

We conclude that \( \phi(\alpha(t)) \) is increasing and strictly concave on this interval.

We can now eliminate from consideration any \( \alpha \) with \( \mu_A/(\mu_A + \mu_B) \leq \alpha \leq \mu_A/\mu_B \).

**Lemma 5 (No Equilibrium for Intermediate \( \alpha \)).** If \( \mu_A/(\mu_A + \mu_B) \leq \alpha(t) \leq \mu_A/\mu_B \), then we have \( \alpha(t+1) = \phi(\alpha(t)) \geq \alpha(t) \). Therefore, this interval does not contain an equilibrium value for \( \alpha \).

**Proof of Lemma 5.** Substituting into equation (9), we have

\[ \phi\left(\frac{\mu_A}{\mu_A + \mu_B}\right) = 1 - \frac{\mu_A^2}{2\mu_B^2}. \]

Then, \( \mu_A < \mu_B \) implies

\[ \mu_A^3 + \mu_B^2 \mu_A < 2\mu_A^3 \]
\[ \iff 2\mu_A^2 \mu_B + \mu_B^2 (\mu_A + \mu_B) < 2\mu_B^2 (\mu_A + \mu_B) \]
\[ \iff \frac{\mu_A}{\mu_A + \mu_B} < 1 - \frac{\mu_A^2}{2\mu_B^2} = \phi\left(\frac{\mu_A}{\mu_A + \mu_B}\right). \]

(21)

Again substituting into equation (9), we get

\[ \phi\left(\frac{\mu_A}{\mu_B}\right) = 1 > \frac{\mu_A}{\mu_B}. \]

(22)

By Lemma 4, the function \( \phi(\alpha(t)) \) is concave for \( \mu_A/(\mu_A + \mu_B) \leq \alpha(t) \leq \mu_A/\mu_B \). By the definition of concavity—see, e.g., Boyd and Vandenberghe (2004, Section 3.1)—for any \( \theta \in [0, 1] \), we then have

\[ \phi(\theta x + (1-\theta)x') \geq \theta \phi(x) + (1-\theta)\phi(x') \quad \text{for} \quad x, x' \in \left[\frac{\mu_A}{\mu_A + \mu_B}, \frac{\mu_A}{\mu_B}\right]. \]

(23)

We can express any \( \alpha(t) \in [\mu_A/(\mu_A + \mu_B), \mu_A/\mu_B] \) as

\[ \alpha(t) = \theta\left(\frac{\mu_A}{\mu_A + \mu_B}\right) + (1-\theta)\left(\frac{\mu_A}{\mu_B}\right), \]

where \( \alpha(t) \) ranges from \( \mu_A/(\mu_A + \mu_B) \) to \( \mu_A/\mu_B \) as \( \theta \) ranges from 0 to 1. Taking \( x = \mu_A/(\mu_A + \mu_B) \) and \( x' = \mu_A/\mu_B \) in equation (23), we get

\[ \phi(\alpha(t)) = \phi\left(\theta\left(\frac{\mu_A}{\mu_A + \mu_B}\right) + (1-\theta)\left(\frac{\mu_A}{\mu_B}\right)\right) \geq \theta \phi\left(\frac{\mu_A}{\mu_A + \mu_B}\right) + (1-\theta)\phi\left(\frac{\mu_A}{\mu_B}\right) \]
\[ > \theta\left(\frac{\mu_A}{\mu_A + \mu_B}\right) + (1-\theta)\left(\frac{\mu_A}{\mu_B}\right) = \alpha(t), \]

where the strict inequality holds because \( \phi(\mu_A/(\mu_A + \mu_B)) > \mu_A/(\mu_A + \mu_B) \) and \( \phi(\mu_A/\mu_B) > \mu_A/\mu_B \), by equations (21) and (22), respectively. We conclude that \( \phi(\alpha(t)) > \alpha(t) \) for all \( \mu_A/(\mu_A + \mu_B) \leq \alpha(t) \leq \mu_A/\mu_B \), and hence that no interior equilibrium can exist in this interval.

\[ \square \]
The only remaining candidates for interior equilibria are any $\alpha$ satisfying $0 < \alpha < \mu_A/\mu_A + \mu_B$. We now establish bounds on the quantities $\alpha_-^*$ and $\alpha_+^*$, and we then show that these are the only possible interior equilibria. Recall from equation (12) that

$$\alpha_-^* = \frac{1}{2} - \frac{\mu_A}{4\mu_B} - \nu,$$

and

$$\alpha_+^* = \frac{1}{2} - \frac{\mu_A}{4\mu_B} + \nu,$$

where

$$\nu = \frac{\sqrt{\mu_A^2 + 12\mu_A\mu_B - 12\mu_B^2}}{4\mu_B}.$$

**Lemma 6 (Bounds on $\alpha_-^*$ and $\alpha_+^*$).** If $\eta < \mu_A/\mu_B < 1$, then we have

$$0 < \alpha_-^* < \alpha_+^* < \frac{\mu_A}{\mu_A + \mu_B}.$$

**Proof of Lemma 6.** If $\eta < \mu_A/\mu_B < 1$, then $\nu$ is real and positive, and clearly then we have $0 < \alpha_-^* < \alpha_+^*$. If we let $\mu_A/\mu_B = x$ and $g(x) = \alpha_-^*$, expressed as a function of the service rate ratio, then we have

$$g(x) = \frac{1}{2} - \frac{1}{4} x - \frac{1}{4} \sqrt{x^2 + 12x - 12}$$

$$> \frac{1}{2} - \frac{1}{4} \sqrt{1 + 12 - 12}$$

$$= 0,$$

and therefore $\alpha_-^* > 0$ for $\eta < \mu_A/\mu_B < 1$.

A bit more care is required to show that $\alpha_+^* < \mu_A/\mu_A + \mu_b$. Let $\mu_A/\mu_B = x$—and therefore $\mu_A/(\mu_A + \mu_B) = x/(x + 1)$. Define $h(x) = \alpha_+^* - x/(x + 1)$. Then we have

$$h(x) = \frac{1}{2} - \frac{1}{4} x + \frac{1}{4} \sqrt{x^2 + 12x - 12} - \frac{x}{x + 1}.$$

Differentiating, we get

$$h'(x) = -\frac{1}{4} + \frac{2x + 12}{8\sqrt{x^2 + 12x - 12}} - \frac{1}{x + 1} + \frac{x}{(x + 1)^2}$$

$$\geq \frac{x + 1}{4} + \frac{x}{(x + 1)^2}$$

$$> 0.$$

The function $h(x)$ is therefore strictly increasing on $(\eta, 1]$. Substituting $x = 1$, we get $h(1) = 0$. The fact that $h(x)$ is strictly increasing then implies that

$$h(x) < 0 \quad \text{for } \eta < x < 1,$$

and hence that

$$\alpha_+^* < \frac{\mu_A}{\mu_A + \mu_B} \quad \text{for } \eta < \frac{\mu_A}{\mu_B} < 1.$$

We conclude that

$$0 < \alpha_-^* < \alpha_+^* < \frac{\mu_A}{\mu_A + \mu_B}.$$
With these lemmas, we can now complete the proof of the proposition.

Proof of Proposition. An equilibrium requires that \( \phi(\alpha) = \alpha \), and we have ruled out as candidates all but the interval \((0, \mu_A/(\mu_A + \mu_B))\). By equation (9), we can express the appropriate fixed point equation for this interval by

\[
\alpha = 1 - \mu_A/\mu_B + \alpha \left(1 - \frac{\mu_A}{2\mu_B}\right).
\]

(Rearranging equation (25) gives

\[
\alpha^2 + \alpha \left(\frac{\mu_A}{2\mu_B} - 1\right) + 1 - \frac{\mu_A}{\mu_B} = 0.
\]

The solutions to this quadratic equation occur at

\[
\alpha^- = \frac{1}{2} - \frac{\mu_A}{4\mu_B} - \nu,
\]

and

\[
\alpha^+ = \frac{1}{2} - \frac{\mu_A}{4\mu_B} + \nu.
\]

Case 1: \( 0 < \mu_A/\mu_B < \eta \). In this case, the inside of the radical in \( \nu \) is negative and both \( \alpha^- \) and \( \alpha^+ \) are imaginary; thus in this regime there are no interior equilibria and \( E = \emptyset \).

Case 2: \( \mu_A/\mu_B = \eta \). If \( \mu_A/\mu_B = \eta \), then \( \nu = 0 \). The fixed point equation (25) thus has a double root at \( \alpha^- = \alpha^+ = 2 - \sqrt{3} \). We have

\[
.27 \approx 2 - \sqrt{3} < \frac{\mu_A}{\mu_A + \mu_B} = \frac{\eta}{\eta + 1} \approx .48.
\]

Therefore, the right-hand side of equation (25) indeed governs the function \( \phi \) at \( \alpha = 2 - \sqrt{3} \), implying that \( \phi(2 - \sqrt{3}) = 2 - \sqrt{3} \). We conclude that \( E = \{2 - \sqrt{3}\} \).

Case 3: \( \eta < \mu_A/\mu_B \). In this case, by Lemma 6, we have

\[
0 < \alpha^- < \alpha^+ < \frac{\mu_A}{\mu_A + \mu_B}.
\]

Thus, both solutions to the fixed point equation fall into the interval in which the expression on the right-hand side of equation (25) governs the function \( \phi \). We conclude that

\[
\phi(\alpha^-) = \alpha^- \quad \text{and} \quad \phi(\alpha^+) = \alpha^+,
\]

i.e., that both \( \alpha^- \) and \( \alpha^+ \) are interior equilibria. Thus, we have \( E = \{\alpha^-, \alpha^+\} \). □

Proof of Proposition 3. We proceed by cases.

Case 1: \( 0 < \alpha^{(1)} < \alpha^- \). By Lemmas 3 and 6, the function \( \phi(\alpha) \) is strictly increasing and strictly convex for \( 0 < \alpha \leq \alpha^- < \mu_A/(\mu_A + \mu_B) \) (equation (27) shows that \( \alpha^- = \alpha^+ < \mu_A/(\mu_A + \mu_B) \) when \( \mu_A/(\mu_A + \mu_B) = \eta \), ensuring that Lemma 3 applies in that case). Moreover, Proposition 2 and the assumption that \( \eta \leq \mu_A/\mu_B \) give us that \( \phi(\alpha^-) = \alpha^- \). These observations together imply that

\[
\phi(\alpha) < \alpha^- \quad \text{for any} \quad 0 < \alpha < \alpha^-.
\]

Thus, the sequence \( \{\alpha^{(i)}\} \) is bounded above by \( \alpha^- \) for any sequence \( \{\alpha^{(i)}\} \) with \( 0 < \alpha^{(1)} < \alpha^- \). Furthermore, consider \( 0 < \epsilon < \alpha^- \). Simple algebra shows that \( \phi(\epsilon) > \epsilon \) if and only if

\[
\epsilon \left(1 - \frac{\mu_A}{2\mu_B}\right) - \epsilon^2 < 1 - \frac{\mu_A}{\mu_B}.
\]
Take $\epsilon < 1 - \mu_A/(2\mu_B - \mu_A)$. Then we have

$$
\epsilon < 1 - \frac{\mu_A}{2\mu_B - \mu_A}
= \epsilon(1 - \frac{\mu_A}{2\mu_B}) < 1 - \frac{\mu_A}{\mu_B}
$$

Thus, there exists $0 < \epsilon < \alpha^*$ with $\phi(\epsilon) > \epsilon$. Proposition 2 guarantees that $\alpha^*$ and $\alpha^+_*$ are the only possible interior solutions to the equation $\phi(\alpha) = \alpha$. Then, because $\phi(\epsilon) > \epsilon$ and $\phi(\alpha) \neq \alpha$ for all $0 < \alpha < \alpha^*$, the continuity of $\phi$ requires that

$$
\alpha^{(t+1)} = \phi(\alpha^{(t)}) > \alpha^{(t)} \quad \text{for } t = 1, 2, \ldots
$$

Hence, in this case the sequence $\{\alpha^{(t)}\}$ is strictly increasing in $t$. Because the sequence is nondecreasing and bounded above, it converges by the monotone convergence theorem (Rudin 1976, Theorem 3.14 on p. 55). To show that it converges to $\alpha^*$ in particular, we suppose otherwise and derive a contradiction. Let $0 < \kappa < \alpha^*$, and suppose that

$$
\lim_{t \to \infty} \alpha^{(t)} = \lim_{t \to \infty} \alpha^{(t+1)} = \lim_{t \to \infty} \phi(\alpha^{(t)}) = \kappa. \quad (28)
$$

We immediately have $\phi(\kappa) \neq \kappa$ because $\kappa \notin \{0, \alpha^*, \alpha^*_+, 1\}$, the set of possible fixed points of the function $\phi$. The function $\phi$ is continuous on $(0, \alpha^*)$. By Lang (1997, Theorem 2.1 on p. 171), for a continuous function $\phi$ and a sequence of real numbers $\{\alpha^{(t)}\}$ that converges to $\kappa$, we must have

$$
\lim_{t \to \infty} \phi(\alpha^{(t)}) = \phi(\kappa),
$$

which when combined with equation (28) contradicts that $\phi(\kappa) \neq \kappa$. We deduce then that $\{\alpha^{(t)}\}$ cannot converge to $\kappa < \alpha^{(t)}$. Because the sequence converges and is bounded above by $\alpha^*$, we conclude that it must converge to $\alpha^*_+$.  

**Case 2:** $\alpha^{(1)} = \alpha^*_+$. In this case, $\phi(\alpha^*_+) = \alpha^*_+$ implies that $\alpha^{(t)} = \alpha^*$ for all $t \geq 1$, so the limit of the sequence is necessarily $\alpha^*_+$. 

**Case 3:** $\alpha^*_- < \alpha^{(1)} < \alpha^*_+$. The proof for this case is similar to that for Case 1. For $\alpha^*_- < \alpha < \alpha^*_+$, let $\alpha = \theta \alpha^*_- + (1 - \theta) \alpha^*_+$ for some $0 < \theta < 1$. By the definition of strict convexity (Boyd and Vandenberghe 2004, Section 3.1)—which holds for $\phi$ in this range of $\alpha$ by Lemmas 3 and 6—we have

$$
\phi(\theta \alpha^*_- + (1 - \theta) \alpha^*_+) < \theta \phi(\alpha^*_-) + (1 - \theta) \phi(\alpha^*_+) = \theta \alpha^*_- + (1 - \theta) \alpha^*_+; \quad (29)
$$

where the equality follows because $\alpha^*_-$ and $\alpha^*_+$ are fixed points of $\phi$ by Proposition 2. Substituting $\alpha = \theta \alpha^*_- + (1 - \theta) \alpha^*_+$ into equation (29) then gives

$$
\phi(\alpha) < \alpha \quad \text{for } \alpha^*_- < \alpha < \alpha^*_+. \quad (30)
$$

Equation (30) implies that the sequence $\{\alpha^{(t)}\}$ is strictly decreasing in $t$. Combined with the fact that $\phi(\alpha^*_-) = \alpha^*_-$, that the function $\phi(\alpha)$ is strictly increasing in $\alpha$ for $0 < \alpha < \alpha^*_+$ by Lemmas 3 and 6 implies that

$$
\phi(\alpha) > \alpha^*_- \quad \text{for } \alpha^*_- < \alpha < \alpha^*_+. 
$$
Thus, the sequence converges because it is nonincreasing and bounded below by \( \alpha^- \) (Rudin 1976, Theorem 3.14 on p. 55).

The sequence cannot converge to \( \alpha^*_+ \) because it is bounded above by \( \alpha^{(1)} < \alpha^*_+ \). Moreover, a similar argument as in Case 1 shows that \( \lim_{t \to \infty} \alpha^{(t)} \neq \kappa \) for any \( \alpha^- < \kappa < \alpha^{(1)} \). Therefore, since the sequence is bounded between \( \alpha^- \) and \( \alpha^{(1)} \) and it does not converge to any \( \kappa \) with \( \alpha^- < \kappa \leq \alpha^{(1)} \), the fact that it indeed converges implies that it must converge to \( \alpha^- \), i.e., we have

\[
\lim_{t \to \infty} \alpha^{(t)} = \alpha^-.
\]

Proof of Proposition 4. We proceed by cases.

Case (i).1: \( \alpha^*_+ < \alpha^{(1)} < 1 \). Proposition 2 implies that

\[
\phi(\alpha) \neq \alpha \quad \text{for any } \alpha^*_+ < \alpha < 1.
\]  

(31)

Furthermore, we know from Lemma 5 that \( \phi(\alpha) > \alpha \) for \( \mu_A/(\mu_A + \mu_B) \leq \alpha \leq \mu_A/\mu_B \). So, there exists \( \alpha \in (\alpha^*_+, 1) \) such that \( \phi(\alpha) > \alpha \). Because \( \phi \) is continuous on \((0, 1] \)—the piecewise sections coincide at their boundaries—the intermediate value theorem applied to the function \( g(\alpha) = \phi(\alpha) - \alpha \) then implies that if there exists \( \alpha' \in (\alpha^*_+, 1) \) with \( \phi(\alpha') < \alpha' \), then there must also exist \( \alpha'' \) in the same interval such that \( \phi(\alpha'') = \alpha'' \).

But this contradicts equation (31), and therefore we conclude that

\[
\phi(\alpha) > \alpha \quad \text{for all } \alpha^*_+ < \alpha < 1.
\]  

(32)

Together with the fact that \( \phi(1) = 1 \) by Proposition 1, equation (32) then implies that if \( \alpha^*_+ < \alpha^{(1)} \), then the sequence \( \{\alpha^{(t)}\} \) is nondecreasing in \( t \). The fact that \( \{\alpha^{(t)}\} \) is bounded above by 1 then gives us that the sequence must converge. A similar argument to that used in the proof of Proposition 3 shows that the sequence must not converge to any \( \kappa < 1 \), and the fact that it converges then implies that it must converge to 1, i.e., that \( \lim_{t \to \infty} \alpha^{(t)} = 1 \).

Case (i).2: \( \alpha^{(1)} = 1 \). By Proposition 1, if \( \alpha^{(1)} = 1 \), then \( \alpha^{(t)} = 1 \) for all \( t \geq 2 \).

Case (ii). If \( \mu_A/\mu_B < \eta \), then by Proposition 2, there are no interior equilibria, i.e., we have \( \phi(\alpha) \neq \alpha \) for all \( 0 < \alpha < 1 \). From equation (9), we know that \( \phi(\mu_A/\mu_B) = 1 > \mu_A/\mu_B \). An application of the intermediate value theorem, similar to that used in part (i), then implies that \( \phi(\alpha) > \alpha \) for all \( 0 < \alpha < 1 \). Moreover, the sequence is again bounded above by 1. Because also \( \phi(1) = 1 \), we then have that \( \{\alpha^{(t)}\} \) is a nondecreasing sequence that is bounded above, which therefore must converge. Similar arguments to those used in the proof of Proposition 3 demonstrate that the sequence cannot converge to any \( \kappa < 1 \). We therefore conclude that the sequence converges to 1, i.e., that \( \lim_{t \to \infty} \alpha^{(t)} = 1 \). 

Proof of Proposition 5. The overall cumulative system time can be broken down into two components: the cumulative system time for \( AB \) customers and the cumulative system time for \( BA \) customers. Respectively denoting these by \( d_A(\cdot) \) and \( d_B(\cdot) \), we have

\[
D(\alpha, \mu_A, \mu_B) = d_A(\alpha, \mu_A, \mu_B) + d_B(\alpha, \mu_A, \mu_B).
\]  

(33)
First, we compute the cumulative system time for $BA$ customers. By Lemma 2, for a customer in position $y_B$, the system time is

$$B(\alpha; y_B) = \frac{\alpha + y_B}{\mu_A}.$$ 

Integrating over all customers, we get that the cumulative system time for $BA$ customers is

$$d_B(\alpha, \mu_A, \mu_B) = \int_0^{1-\alpha} \frac{\alpha + y_B}{\mu_A} dy_B = \frac{\alpha(1-\alpha)}{\mu_A} + \frac{(1-\alpha)^2}{2\mu_A}.$$ 

For $AB$ customers, by Lemma 1, for a customer in position $y_A$, the system time is

$$A(\alpha; y_A) = \begin{cases} \frac{y_A + 1 - \alpha}{\mu_B} & y_A \leq \mu_A \left( \frac{1 - \alpha}{\mu_B - \mu_A} \right), \\ \frac{y_A}{\mu_A} & \text{otherwise}. \end{cases}$$

Recall from equation (13) that

$$\xi_\alpha = \min \left\{ \alpha, \mu_A \left( \frac{1 - \alpha}{\mu_B - \mu_A} \right) \right\}.$$ 

Integrating, we get

$$d_A(\alpha, \mu_A, \mu_B) = \int_0^\alpha A(\alpha; y_A) dy_A = \int_0^{\xi_\alpha} \frac{y_A + 1 - \alpha}{\mu_B} dy_A + \int_{\xi_\alpha}^\alpha \frac{y_A}{\mu_A} dy_A = \frac{\xi_\alpha(1-\alpha)}{\mu_B} + \frac{\xi_\alpha^2}{2\mu_B} + \frac{\alpha^2}{2\mu_A} - \frac{\xi_\alpha^2}{2\mu_A}.$$ 

Note that if $\xi_\alpha = \alpha$, then the range of the second integral above is empty (coinciding with the latter two terms of the reduced expression canceling each other out), and we therefore conclude

$$d_A(\alpha, \mu_A, \mu_B) = \begin{cases} \frac{\alpha(1-\alpha)}{\mu_B} + \frac{\alpha^2}{2\mu_B} & \text{if } \alpha \leq \frac{\mu_A}{\mu_B}, \\ \frac{\xi_\alpha(1-\alpha)}{\mu_B} + \frac{\xi_\alpha^2}{2\mu_B} + \frac{\alpha^2}{2\mu_A} - \frac{\xi_\alpha^2}{2\mu_A} & \text{otherwise}. \end{cases}$$

Substituting our expressions for $d_A(\cdot)$ and $d_B(\cdot)$ into equation (33) gives the formula in equation (14). \hfill \Box

References


