Service Networks with Open Routing and Procedurally Rational Customers

Andrew E. Frazelle
Jindal School of Management, The University of Texas at Dallas, Richardson, TX 75080
andrew.frazelle@utdallas.edu

Tingliang Huang
Haslam College of Business, The University of Tennessee, Knoxville, Knoxville, TN 37996
thuang7@utk.edu

Yehua Wei
The Fuqua School of Business, Duke University, Durham, NC 27708
yehua.wei@duke.edu

Self-interested customers populate various service systems, and these customers may not be fully rational. Customers' form of reasoning and its consequences for system performance affect the planning decisions of service providers. We study *procedurally rational* customers—that is, customers who make decisions based on a sample containing anecdotes of the system times experienced by other customers. Specifically, we investigate the implications of procedurally rational customers on service networks with open routing, i.e., those in which customers visit multiple stations but can choose the order in which to visit the stations. Because some actions may be less represented in the population, a given customer may not be successful in obtaining anecdotes about all possible actions. We introduce a novel sampling process that extends the procedural rationality framework to incorporate the *discernibility* of customers: better discernibility implies that customers are more likely to obtain anecdotes about all actions. We characterize the response of procedurally rational customers under this model. We study equilibrium routing profiles, where the fraction of customers choosing each route becomes stationary. As the sample size grows large, customers' estimates become more accurate, and the procedurally rational equilibrium converges to the fully rational equilibrium (which is also socially optimal). More strikingly, the procedurally rational outcome also converges to the fully rational equilibrium as the discernibility parameter grows large, *even if the number of anecdotes remains small.* To achieve a good customer experience, it is crucial for customers to obtain anecdotes about each alternative. In our open-routing service network, if procedurally rational customers can obtain anecdotes about all actions (high discernibility), then despite the sampling error, their decisions will closely resemble those of fully rational customers. If they cannot (low discernibility), then their choices can deviate significantly from those of fully rational customers, leading to substantial excess waits.

Key words: open routing, bounded rationality, anecdotal reasoning, behavioral operations, queueing
This version: August 24, 2022.

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, Cui et al. 2019); queueing networks (Cohen and Kelly 1990, Honnappa and Jain 2015); restaurant reservations (Çil and Lariviere 2013, Oh and Su 2018); on-demand healthcare platforms (Liu et al. 2018); organ donation and transplantation (Su and Zenios 2004, Dai et al. 2020, Nageswaran and Scheller-Wolf 2022); 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 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. Additionally, Honnappa and Jain (2015) study the "network concert queueing game," which involves customers choosing their arrival times to a queueing network as well as their routes through the network.

The existing literature on strategic open routing has assumed customers to be fully rational. However, an open-routing service network is a complex system, and even to compute (much less implement) the fully rational equilibrium requires customers to know all system parameters and then map out the interaction of multiple queues over all possible collective routing decisions. Even if some customers are sophisticated, the question remains what each customer assumes about the rationality of the others. Therefore, it is perhaps likely that customers will not behave like fully rational agents in such systems. Instead, they may adopt simple heuristics to decide on their routes, exhibiting bounded rationality. Bounded rationality has been studied in several strategic queueing contexts in recent years (see, e.g., Huang et al. 2013 and other references in Section 2), mostly in queueing systems with only one station. However, we are not aware of any work that studies bounded rationality in a multi-station, open-routing service network, which we believe to be the type of nuanced setting in which customers are likely to exhibit such behavior.

The present work aims to address this gap. Relaxing the assumption of full rationality opens up a world of possible modeling choices to incorporate customer bounded rationality. Some common choices in the behavioral operations literature include quantal response equilibrium (Su 2008), reference dependence (Wu et al. 2015, Yang et al. 2018), and procedural rationality (Ren et al. 2018). Modeling bounded rationality is complex in almost any context. In this paper, we propose a tractable model of bounded rationality that extends the framework of procedural rationality. The notion of procedural rationality was introduced by Osborne and Rubinstein (1998) as an alternative to the traditional Nash equilibrium concept traditionally used for noncooperative games. Their $S(K)$-reasoning framework (K representing a sample size) 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 addition to Osborne and Rubinstein (1998), he also motivates his model of anecdotal reasoning from previous work (Tversky and Kahneman 1971) that reported experiments in which human decision makers "over-infer from small samples. They explained this phenomenon (dubbed 'the law of small numbers') as a consequence of the 'representativeness' heuristic: people expect a small sample to mirror the probability distribution from which it is drawn" (Spiegler 2006, p. 1115). The so-called "sample naivete" phenomenon is also documented and studied in the psychology literature; see, e.g., Fiedler (2000), Fiedler et al. (2006), Juslin et al. (2007), and references therein. Finally, in an operational context, Tong and Feiler (2017) adapt these concepts from psychology to study the "naive intuitive statistician," which is closely related to procedural rationality but does not involve reasoning about others' actions. In their model, a planner draws a "mental sample" of finite size from a probability distribution and mistakenly treats the properties of that sample as those of the true distribution.

The framework of procedural rationality has the benefit of requiring only a single parameter: the number of "anecdotes" that each player takes for a given action. Moreover, it emulates a familiar formula from daily life of reasoning via anecdotes: asking colleagues to recommend a doctor after moving to a new city; asking a friend about her experience at a newly opened restaurant; or sampling different routes on a daily commute to assess which is best. For this reason, procedural rationality is also called *anecdotal reasoning* because individuals reason based on small samples or anecdotes, which constitute "word of mouth." Indeed, word of mouth is an important factor in consumer purchasing decisions, particularly when a wait is involved: in-store marketing firm Spectrio states about managing customer queues that "A good perception of their [the customer's] experience means they'll consider becoming repeat customers and possibly pass on word-of-mouth advertising to their friends and family" (Spectrio 2015).

Aside from simplicity and its resemblance to human behavior, another important reason for choosing the procedural rationality framework is that the base model naturally extends to incorporate several important features in the setting considered in this paper. In our open routing setting, the anecdotes are categorized by routes through a service network and there are multiple routes to evaluate and compare. However, it is natural to expect that anecdotes associated with a certain route that is less represented in the population may be harder to sample than others, a feature that is not captured by existing models of procedural rationality. For concreteness, consider the following real-world example. A guest planning her trip to Disney World might post on social media, asking about others' experiences with different routes through the park. A recent such post from July 2021, in the Facebook Group "Walt Disney World Tips and Tricks"[1] (a group with over 600,000 members as of August 2022) reads "We're headed to WDW... our first day we will be at Magic Kingdom... We plan to [arrive at] rope drop, what route should we take at the park...?"[2] The post received several replies with suggested routes. "Rope drop" occurs at the opening of the park, when a large number of guests enter and make their routing decisions around the same time (this resembles our model setting with customers present at the start of service). For simplicity, let us consider a hypothetical guest who makes a similar post, in a simplified system with only two rides (say, A and B): "which ride did you visit first, A or B, and how long were your wait times at both rides?" The replies to this post can be assumed to depend on the population proportion of routes, and if one of the routes is rare in the population, then it may not appear at all in the comments. Suppose that the only route reflected is the route visiting A first and then B. Having only received anecdotes about one route, the guest may post again, asking specifically "did anyone go to B first and then A? What were your wait times?" If the missing route is rare in the population, then she may not receive replies because of the scarcity of that route. In this case, she may choose her route arbitrarily. However, the activeness and size of the group and the willingness of the users to help others both moderate the probability that her second post (about the missing route) will receive replies. For a given population proportion of routes, she will be more likely to receive a reply about the missing route if the group is larger and more active.

In the procedural rationality framework, and motivated by practical examples of customer reasoning about routing decisions like the one above, we seek a customer sampling model with four important facets, namely one that: (i) is parsimonious (ideally, one or few parameters); (ii) captures the fact that the probability of obtaining both types of anecdotes is significantly affected by the population proportion; (iii) is reasonably aligned with customer behavior; and (iv) allows

[1] https://www.facebook.com/groups/disneytipsandtricks

[2] https://www.facebook.com/groups/disneytipsandtricks/posts/1930298137135681

the probability of obtaining both types of anecdotes to differ from the product of population proportions. To this end, we introduce a sequential sampling process with a *discernibility* parameter that extends the $S(K)$-reasoning framework to model customer behavior. Roughly speaking, the discernibility parameter reflects the likelihood of a player successfully obtaining a diverse set of anecdotes in a small sample, and it is determined by multiple factors including players' effort, system characteristics and service provider intervention. To the best of our knowledge, the sequential sampling process (with discernibility) we propose is new, and can be therefore of independent interest to the bounded rationality literature. In the open routing setting we study, the inclusion of the discernibility parameter not only improves the fidelity of the model, but also uncovers the following important insight: the discernibility, and therefore the diversity of anecdotes, can drive the procedurally rational equilibrium to the outcome of a fully rational model (Nash equilibrium), even when each player (customer) draws only a small sample.

For the open-routing network, 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 exhibit procedural rationality in their reasoning about wait times. The routing game is played by the customers, with each customer making a decision based on the waiting time experienced by several other customers she sampled from. While we model it as a static game, it could be also viewed as a repeated game where the customers from the current period sample randomly from the past. To understand the equilibrium of customers' behavior, we study a fluid approximation with a continuum of infinitesimal customers. In the fluid approximation, we characterize the response function of procedurally rational customers. This characterization enables us to demonstrate the equilibrium of customers' behavior as the number of anecdotes and discernibility changes, and to fully characterize the equilibrium when the number of anecdotes or the discernibility parameter approaches infinity. 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 a repeated game setting.

We study the difference between the equilibrium outcomes under procedural rationality and those with fully rational customers. In the fully rational model, Arlotto et al. (2019) observe that customers *herd* by all following the same route through the network; by contrast, we find that, in the open-routing setting, procedural rationality does not produce a herding outcome. This finding is similar to studies in the literature, where many have observed outcomes under procedural rationality that differ markedly from the outcomes with fully rational customers (see Section 2). A more striking finding, however, is that the outcome with procedurally rational customers converges to herding as the discernibility parameter approaches infinity, *even when the number of anecdotes is small.*

This suggests that, as long as the customers obtain anecdotes of different types, the equilibrium outcome with procedurally rational customers would be similar to that with fully rational customers, even though the equilibria for the two models arise from fundamentally different decision-making processes. In addition, we also find that herding produces the socially optimal outcome for the customers. As a result, our findings provide the following important insight to service providers in the open-routing setting: to achieve a good customer experience, it is crucial that customers obtain anecdotes about *each alternative*. This finding reflects both the robustness of herding and the vulnerability of customers who reason based on anecdotes without underlying knowledge of the system. To wit, if procedurally rational customers receive even a little information about each route, then they approximate the rational (and socially optimal) behavior of herding, but if instead one of the routes is missing from their sample, then they have no basis for comparison and can stray arbitrarily far from herding, which can result in significant excess waiting times.

The remainder of the paper is organized as follows. Section 2 reviews the related literature, and Section 3 describes the routing decisions of procedurally rational customers and a fluid approximation for customer routing. The routing decisions of both procedurally and fully rational customers are derived and compared in Section 4. Section 5 studies the performance of the system for different equilibrium outcomes. Section 6 conducts numerical studies, and Section 7 concludes the paper. All proofs can be found in the online appendix.

2. Related Literature

As mentioned, Arlotto et al. (2019) find that in an open-routing service network, fully rational customers herd, and one of our key findings is that procedurally rational customers in such a system also herd if discernibility is high. Herding behavior has also been observed in the economics literature (see Smith and Sørensen 2020 for a recent example) as well as in other queueing-related settings (see Kremer and Debo 2016, Veeraraghavan and Debo 2011, among others). In Smith and Sørensen (2020), customers are Bayesian and use the actions taken by previous customers to update their beliefs about the utility of different actions. If a string of customers chooses the same action, this influences the later customers to increase their quality belief for that action, which can lead to herding. In prior queueing studies including the two mentioned above, the reason for herding is that when customers choose between service providers and some have private signals about quality, the queue length conveys information about the quality of a service provider; this can lead to customers joining a longer queue to obtain higher quality service because the difference in quality can outweigh the increased waiting cost. Crucially, in all studies of herding behavior that we are aware of—apart from the open routing setting, that is—the driver of herding is *informational*. By contrast, in an open routing setting, herding is *strategic*. The more customers that choose a given route, the better

that route becomes relative to other routes (see Arlotto et al. 2019); this is strikingly different from herding in other contexts, in which making the same decision as others either has no direct impact on utility (Smith and Sørensen 2020 and earlier studies of informational herding in economics) or actually *harms* the utility conditional on the quality level because it increases waiting time (Kremer and Debo 2016, Veeraraghavan and Debo 2011). Lastly, in the procedural rationality setting of this paper, customers reason based on samples. While this might appear similar to reasoning based on the actions of others leading to informational herding, it is in fact fundamentally different. First, procedurally rational customers are not Bayesian but rather reason heuristically, and second, they decide based not only on the *actions* observed in their sample (as in the studies mentioned above) but also on the *consequences* of those actions, i.e., the realized system times.

More broadly, 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 Yang and Debo (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 behavioral 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 recent 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; Debo and Snitkovsky (2018) on tipping and social norms; and Yang et al. (2018) on loss-averse customers.

Within the operations management literature, several recent papers study procedurally rational customers, e.g., Huang and Yu (2014) on opaque selling, 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 customers using anecdotal reasoning to infer *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. Another significant difference between this study and the existing economics and operations literature is that we have two types of anecdotes coming from two stochastic systems/queues, whereas the literature frequently considers only a single type of anecdotes. This difference is nontrivial because in our setting, we must consider how customers collect those different types of anecdotes depending on their efforts as well as the system characteristics, and how to meaningfully utilize them. We propose a novel model to address these critical issues.

To the best of our knowledge, the only previous work to assume that customers use anecdotal reasoning to infer waiting time in a service 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. They contrast customers who reason via anecdotes with their fully rational counterparts and show important differences in the optimal pricing strategy with procedurally rational (those who reason via anecdotes) versus fully rational customers. By contrast, we demonstrate the nuanced role that procedural rationality plays in an open-routing service network, and as mentioned above, our context involves customers receiving anecdotes about multiple alternatives. In a service network with routing decisions, procedurally rational customers may behave either approximately the same as or completely differently from fully rational customers, depending largely on customers' ability to obtain anecdotes about multiple routes. In addition, we propose a sampling model to handle the possibility that some customers will fail to obtain all types of anecdotes, which is novel in the procedural rationality literature.

Outside the operations literature, some existing sampling methods from the machine learning, statistics, and marketing literatures have considered forms of biased sampling. However, as we discuss next, these methods, while related, are qualitatively different from our sampling with discernibility, both in aim and in implementation.

We first discuss the method of learning from imbalanced data (see, e.g., He and Garcia 2009). A common machine learning task is to train a classifier to predict the category label of an unlabeled instance. However, to train the classifier well may require over- or under-sampling the under- and over-represented classes, respectively, as otherwise the classifier may not learn enough about under-represented classes; learning from imbalanced data tackles this task in a structured way. Though this approach is related to our model, the two are fundamentally different. In learning from imbalanced data, the training examples are labeled, and the algorithm can select which classes to sample to achieve a certain performance metric; by contrast, our model of sampling with discernibility reflects customers with no explicit knowledge of the data-generating process, and it abstractly captures the exogenous and endogenous factors that influence the probability of obtaining diverse anecdotes. Learning from imbalanced data is also similar to endogenous selective sampling from the marketing literature (see, e.g., Donkers et al. 2003). Selective sampling seeks to learn about customers who choose alternatives that are less common; for example, a company might target a special survey specifically for such customers. Importantly, in this context the marketer wishes to estimate a model of customer behavior but does not himself face the decision that customers in his data faced, unlike in our model in which the decision maker faces the same decision as those she samples from.

Another related concept, this one from statistics, is importance sampling (see, e.g., Owen and Zhou 2000). Importance sampling aims to estimate a specified function of some random variable. For a function that has nonnegligible magnitude only for a low-measure region in the variable's support, importance sampling samples the random variable from a sampling density that differs from the true density, in order to obtain more information about the region that is "important" for the value of the function (i.e., where it is nonnegligible). It then applies a simple correction factor to correct for the biased sampling density. Like learning from imbalanced data, importance sampling is a statistical procedure that requires knowledge of the data-generating process. Furthermore, importance sampling makes inference about only one random quantity and chooses a sampling distribution to accomplish this goal; by contrast, in our model customers wish to compare two or more different random quantities, and discernibility moderates *which* of these random quantities will appear in their sample and how often. Additionally, the goal of biased sampling in this and the other models above is to go from *some but not enough* data about an alternative (or class/category) to *enough* data. By contrast, in our setting discernibility reflects the need to go from *no* data about an alternative (route) to *some* data.

To summarize, sampling models abound in the marketing and statistics/machine learning literatures to account for data that is imbalanced or missing important information. However, these models usually involve calibrated statistical procedures to achieve desired performance measures, which may require larger sample sizes, and moreover a subject matter expert typically conducts the analysis. By contrast, we model customers who are aware that they need information about multiple alternatives (routes), but (i) they typically seek only a few anecdotes rather than a large data set, (ii) they are not statisticians or marketers, and (iii) they reason heuristically, not statistically, about the information they receive. Our sampling procedure with discernibility is designed to model such customers in a tractable way. It simultaneously reflects both customers' bounded rationality and their recognition of the fact—one that should be apparent even to a boundedly rational customer—that a meaningful comparison requires at least *some* information about each alternative. Discernibility also parameterizes their ability to obtain such diverse information, and we show that this feature plays a critical role in determining outcomes.

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 S (Slow) and station F (Fast). Customers can choose whether to visit station S first or station F first, and every customer must visit both stations. A customer who visits station S (F) first is said to have chosen route SF (FS). The service rate at station S (μ_S) is assumed to be less than the service rate at station F (μ_F). Upon completing service at

the first station on her chosen route, a customer immediately joins the back of the queue at the other station. After customers make their routing decisions, customers who chose route SF are sequenced uniformly at random to determine the queueing order at station S, and similarly for FS customers—that is, those who chose route FS—at station F.

We consider a static routing game where players/customers choose route SF or FS. Our game setting with an open-routing service network resembles that of Arlotto et al. (2019). They consider a standard model with fully rational customers, where each customer optimizes based on the inferred equilibrium strategy of all others as well as the system parameters. By contrast, we propose a model with procedurally rational customers, where each customer does not know the system parameters or consider the strategy of other customers. Instead, customers rely on anecdotes sampled from the experiences of others. The decision process made by the procedurally rational customers can be divided into two components: the process for customers to obtain a sample of anecdotes, and the decision after the sample is obtained.

3.1. Behavior of Procedurally Rational Customers

The decision after collecting/obtaining the anecdotes is similar to that from the literature on procedural rationality (see e.g., Osborne and Rubinstein 1998, Spiegler 2006), with the key novelty that anecdotes can come from either route. As a result, it is possible that a customer obtains all anecdotes from only one route, and to account for this possibility, we assume that customers have a *prior belief* on the expected system time for each route *without any anecdote*. We assume that the priors are extremely noisy (for concreteness, think of a Gaussian prior with standard deviation going to infinity), and hence, the prior about each route is wiped out if any anecdote of the same route is obtained. Therefore, in the event that the sample contains both type FS and SF anecdotes, the customer chooses the route with the lower average system time based on the anecdotes, which is similar to the $S(1)$-reasoning procedure used in Osborne and Rubinstein (1998) and Spiegler (2006). In the event that the sample only contains one type of anecdote, because we assume that the prior on the expected system time is extremely noisy, the customer selects one of the routes SF and FS with equal probability $1/2$ (although our main finding is robust to relaxation of this assumption: see the discussion following Theorem 1).

We next discuss the process for customers to obtain anecdotes (of different routes). To the best of our knowledge, this process has not been considered in the traditional procedural rationality literature, as the literature has mainly focused on systems with customers deciding between an alternative with uncertainty based on anecdotes, and an alternative (e.g., outside option) with a known or deterministic outcome. However, in our setting, a customer will use anecdotes from two

alternatives (routes FS and SF), and in addition, the success rates of obtaining anecdotes from FS and SF should depend on the relative proportion of FS and SF customers in the population.

We model the process for customers to obtain their anecdotes as follows. A customer indexed by i will draw a sample of anecdotes of size K with replacement from the rest of the customers. While it is natural to assume that the sampled customers are uniformly distributed among all customers not indexed by i, we also note that for the information to have value, customer i needs anecdotes from *both* routes. Therefore, in our model, the customer will sample in a way that biases toward obtaining at least one anecdote from each route. Specifically, letting α denote the fraction of SF customers that are not indexed by i, we assume that the first anecdote i draws is selected uniformly at random, and thus, the probability of drawing a type SF anecdote first is α. On any draw after the first, if one type of anecdotes is missing, then the next anecdote is biased towards the missing type. To model the bias, we introduce a *discernibility* parameter $\beta \geq 1$. This parameter can be viewed as a simplified abstraction of the factors other than the population proportion of routes that influence the probability of obtaining the missing type of anecdote, e.g., the size and activeness of the online community in the social media example above. We then let the probability that a type FS anecdote is drawn given that the current sample does not contain type FS be

$$\frac{\beta(1-\alpha)}{\alpha+\beta(1-\alpha)}. \tag{1}$$

Similarly, given that the current sample does not contain type SF, we let the probability that a type SF anecdote is drawn be

$$\frac{\beta\alpha}{\beta\alpha+(1-\alpha)}. \tag{2}$$

When $\beta = 1$, customers do not sample in a way that is biased towards the missing anecdote, while as $\beta \to \infty$, we approach perfect discernibility, as the probability of obtaining the anecdote of the missing type converges to 1 in the second draw. In Appendix A, we provide additional interpretation of the discernibility parameter, including its relationship to the logit choice model.

After obtaining both type SF and FS anecdotes, the customer has information about both routes, and her sampling process may change for future draws. It seems unlikely that, after obtaining the missing type so that a meaningful comparison is possible, her sampling process would become *more* biased. However, her sampling process could be *less* biased toward the less-represented type than when that type was completely absent. To reflect this, we make the following assumption, under which our results hold for the general model, independent of the specific sampling process after both types of anecdotes are obtained.

ASSUMPTION 1. *For each of the K draws, the probability that the anecdote is of type SF (FS) is bounded below by $\underline{\rho}_S$ ($\underline{\rho}_F$), where*

$$\underline{\rho}_S := 1 - \frac{\beta(1-\alpha)}{\alpha + \beta(1-\alpha)} \quad \textit{and} \quad \underline{\rho}_F := 1 - \frac{\beta\alpha}{\beta\alpha + 1 - \alpha}.$$

First, we note that our specification above based on equations (1) and (2) indeed satisfies this assumption for each draw before both types of anecdotes are obtained: the probability of an SF anecdote is either (i) α, if it is the first draw in the sample; (ii) $\beta\alpha/(\beta\alpha + (1-\alpha))$, if we have one or more FS anecdotes but no SF anecdotes; or (iii) $\underline{\rho}_S$, if we have one or more SF anecdotes but no FS anecdotes. Because $\beta \geq 1$, the smallest of these quantities is $\underline{\rho}_S$. Analogous reasoning implies that for each draw before obtaining both types of anecdotes, the probability of an FS anecdote is at least $\underline{\rho}_F$. Accordingly, Assumption 1 merely ensures that after both types of anecdotes are in the sample, each type's probability for future draws falls between the discernibility-biased extremes (e.g., the probability of drawing an SF anecdote will be at least $\underline{\rho}_S$ and at most $1 - \underline{\rho}_F$). Assumption 1 applies for the remainder of the paper.

3.2. Fluid Approximation

For tractability, we consider the game with customers being represented as fluid, which is effectively an approximation of the discrete model when the number of players is large. Fluid-type models have been commonly used in both rational queueing games (Hassin 2016) and the procedural rationality framework (Spiegler 2006), as they simultaneously smooth out the discreteness and allow each player to ignore the impact of her action on others. Later, in Section 6.1, we will conduct numerical studies to ascertain how well the insights from our fluid model carry over into the discrete setting.

In our fluid model, a certain volume (normalized to 1) of fluid must be processed, and the entire volume must be processed at both stations. Station S processes fluid at rate μ_S, and station F processes fluid at rate $\mu_F \geq \mu_S$. Technical details on the system evolution and customer system times in the fluid approximation can be found in Appendix B.

4. Equilibrium Analysis

In this section, we study the equilibrium under our procedurally rational model with different values of the sample size (K) and discernibility parameter (β), and compare it with the equilibrium under the fully rational model. We first show that the decisions of procedurally rational customers will converge to those of fully rational customers as the sample size grows, establishing the connection between procedural and full rationality. After this, we state the central result of our paper, Theorem 1. With this result, we show that for *any* fixed sample size K, as the discernibility improves, the largest procedurally rational equilibrium converges to herding, i.e., fully rational behavior.

We let $\pi_{K,\beta}(\alpha)$ denote the response function with sample size K and discernibility β when a fraction α of customers choose route SF. Specifically, $\pi_{K,\beta}(\alpha)$ is the fraction of customers that will choose route SF, given that all customers follow the procedurally rational behavior described in Section 3.1 with parameters K and β. The response function $\pi_{K,\beta}$ can be characterized as follows. Let $\mathcal{B}$ be the event that a sample collected by a customer contains anecdotes of both SF and FS types, and $\gamma_{K,\beta}(\alpha)$ be its probability. Furthermore, conditioned on $\mathcal{B}$, let $\bar{s}_{K,\beta}(\alpha)$ and $\bar{f}_{K,\beta}(\alpha)$ denote the average system times for routes SF and FS over the anecdotes. Recall that a customer chooses SF with probability $1/2$ if her sample does not contain anecdotes of both types; thus, $\pi_{K,\beta}(\alpha)$ can be expressed as

$$\begin{aligned}\pi_{K,\beta}(\alpha) &= \mathsf{Pr}[\bar{s}_{K,\beta}(\alpha) \le \bar{f}_{K,\beta}(\alpha)|\mathcal{B}] \cdot \mathsf{Pr}[\mathcal{B}] + \frac{1}{2}(1 - \mathsf{Pr}[\mathcal{B}]) \\ &= \mathsf{Pr}[\bar{s}_{K,\beta}(\alpha) \le \bar{f}_{K,\beta}(\alpha)|\mathcal{B}] \cdot \gamma_{K,\beta}(\alpha) + \frac{1}{2}(1 - \gamma_{K,\beta}(\alpha)),\end{aligned} \tag{3}$$

for each $\alpha \in [0,1]$. For any fixed K and β, we say that α is a *procedurally rational equilibrium* whenever $\pi_{K,\beta}(\alpha) = \alpha$.

4.1. Equilibria of Procedurally and Fully Rational Models

To compare the procedurally rational equilibrium with the fully rational equilibrium, we introduce some additional notations. Let $\pi^*(\alpha)$ denote the (fully rational) best response function, i.e., the fraction of customers that will choose route SF by playing their best response on their expected system time for both routes given that a fraction α of customers are choosing SF.[3] As a result, α^* is a Nash equilibrium of the fully rational model if and only if $\pi^*(\alpha^*) = \alpha^*$. Arlotto et al. (2019) proved that all customers herding on route SF (or FS when $2\mu_S > \mu_F$) is a Nash equilibrium in the routing game with discrete customers. The analogous result also holds with a continuum of customers, as the following claim demonstrates.

Claim 1. *For any μ_S and μ_F, we have $\pi^*(1) = 1$. If $2\mu_S \ge \mu_F$, then we also have that $\pi^*(0) = 0$.*

Remark 1. In general, when $2\mu_S \ge \mu_F$, the fully rational model may have a third Nash equilibrium in addition to the two herding equilibria. Throughout, we will focus on the herding equilibrium at SF ($\alpha^* = 1$) for two reasons. First, herding on SF is a Nash equilibrium for all $\mu_S \le \mu_F$, whereas the other Nash equilibria only hold when $\mu_S \le \mu_F \le 2\mu_S$. Second, among all Nash equilibria, the herding equilibrium at SF has the smallest cumulative system time over all customers (see Section 5). Thus, it is in everyone's best interest to coordinate to the herding equilibrium at SF.

[3] When there is a tie in expected system time between SF and FS given that a fraction $0 < \alpha < 1$ of customers are choosing SF, the best response can be either SF or FS. In that case, we can assume that customers in response pick route SF with probability α, which makes α a Nash equilibrium.

We next turn our attention to the equilibrium in our procedurally rational model. In order to provide meaningful comparisons with the Nash equilibrium with $\alpha^* = 1$, we focus on the largest equilibrium in our procedurally rational model Note that a largest equilibrium—which is also the unique equilibrium for a wide parameter range, as discussed after Proposition 3—exists because the set of equilibria is non-empty and compact by Lemma 4 in Appendix D. We now define $\alpha_{K,\beta}$ as the largest procedurally rational equilibrium with fixed K and β.

DEFINITION 1. For any fixed K and β, we define $\alpha_{K,\beta} := \max\{\alpha : \pi_{K,\beta}(\alpha) = \alpha\}$.

One way to compare the outcomes of the procedurally and fully rational models is by looking at the difference between the herding Nash equilibrium ($\alpha^* = 1$) and $\alpha_{K,\beta}$. Intuitively, for each $0 < \alpha < 1$, as the number of anecdotes collected by each customer (denoted by K) becomes large, each customer should get close to a perfect inference on the true expected system time from choosing route SF or FS, given that α fraction of customers are choosing SF. Therefore, the outcome of our procedurally rational model should approach that of the rational model for large K. Our next proposition verifies this intuition, by showing that the largest equilibrium outcome of our model converges to the herding (Nash) equilibrium as K approaches ∞.

PROPOSITION 1 **(Largest Equilibrium Converges to 1 For Large Sample Size)**. *For any fixed $\beta \geq 1$, we have $\lim_{K\to\infty} \alpha_{K,\beta} = 1$.*

In addition to having the same equilibrium outcomes, we note that the response function under our procedurally rational model converges to the best response function under the fully rational model in the interval $(1/2, 1)$ as K approaches infinity. This is stated as Corollary 1, which follows from the proof of Proposition 1.

COROLLARY 1. *For any fixed $\beta \geq 1$, and $\alpha \in (1/2, 1)$, we have $\lim_{K\to\infty} \pi_{K,\beta}(\alpha) = \pi^*(\alpha) = 1$.*

So, for any discernibility, as customers receive more anecdotes, their estimates of the system time approach the true mean; while not altogether surprising, this finding is valuable in that it establishes the connection between procedural rationality and full rationality.

While we have seen that customer behavior under procedural rationality converges to fully rational behavior as the sample size grows, in practice, access to a large number of anecdotes can come at a cost to people, as they may be simply overwhelmed by information or have limited memory (see, e.g., Tong and Feiler 2017 for more on cognitive limitations related to sampling). Therefore, it is important to understand the procedurally rational model with a fixed, finite sample size K. Indeed, the fixed sample size provides the setting for the main result in our paper. A priori,

we might expect that sampling error would cause customers who reason via anecdotes to deviate in their behavior from the fully rational norm of herding, with this deviation being more prominent the smaller the sample size. However, in stark contrast to this expectation, we now show that for *any* non-trivial sample size $K \geq 2$, the largest equilibrium converges to herding (i.e., to $\alpha = 1$) as discernibility increases.

THEOREM 1 **(Convergence to Herding for Any Sample Size)**. *If $\mu_S < \mu_F$, then for any sample size $K \geq 2$: (i) the largest equilibrium $\alpha_{K,\beta}$ converges to herding on route SF as $\beta \to \infty$, i.e., we have $\lim_{\beta\to\infty} \alpha_{K,\beta} = 1$, and (ii) there exists $\hat{\beta}$ such that $\alpha_{K,\beta}$ is increasing in β for all $\beta > \hat{\beta}$.*

In other words, whether the sample size is small or large, as discernibility increases, customer decisions come to resemble those of fully rational customers. The proof of this result hinges on showing that as discernibility increases, the probability of obtaining at least one anecdote of each type converges to 1; indeed, as the proof indicates, a single anecdote of each type is enough to ensure herding behavior. In the extreme case of $K = 2$, each customer is receiving only a small amount of information (at most two anecdotes) from each route type; still, if discernibility is high enough, customers glean enough information about the system times to approximate herding. This finding implies the following important and encouraging insight for managers: increasing the number of anecdotes—which can be thought of as customers becoming more sophisticated in their reasoning—is not necessary to achieve a good customer experience, as long as discernibility is high.

We have assumed that if one type of anecdote is missing from a customer's sample, then she randomizes with equal probability between the routes. However, Theorem 1 is robust to other reasonable assumptions about this case. For instance, it might be supposed that a customer who receives anecdotes from only one route will be more likely to choose that route than the route that she is missing. Under our sampling process, there is a higher probability that a customer will obtain all anecdotes from the more-prevalent route than all from the less-prevalent route. Thus, if customers are more inclined toward the route whose anecdotes they obtain when the other route is missing from their sample, more such customers will choose the more-prevalent route than the less-prevalent route. In short, customers will be pushed more towards herding, further reinforcing the conclusion of Theorem 1. Proposition 5 in Appendix G formalizes this intuition and confirms that the largest equilibrium still converges to herding after relaxing the equal probability assumption.

Our analysis shows that for large α, the supports for the random system time draws on routes SF and FS separate, so that any system time draw from route SF will be shorter than any draw from route FS. We highlight that under procedural rationality, there is a limit on the amount of noise in a customer's estimate of the expected system time. Since her estimate is a sample average

of realizations from the true distribution, it cannot fall outside of the support. This "bound" on the degree of noise is enough to ensure herding; however, under other forms of rationality (e.g., quantal response equilibrium) in which customer's estimates are not bounded by the support of the true distribution, herding would not necessarily arise even if customers were to receive (noisy) information about both routes.

4.2. Procedural Rationality with Small Samples

We have shown that, for any sample size $K \geq 2$, procedurally rational behavior converges to fully rational behavior as discernibility improves. To enrich this finding, and to facilitate sharper analysis, we now study the two-anecdote case in detail for different values of β.

For notational simplicity, we will temporarily drop K from $\pi_{K,\beta}$ and $\gamma_{K,\beta}$, as we will focus on $\pi_{2,\beta}$ and $\gamma_{2,\beta}$ in this subsection. From equations (1) and (2), the probability that a customer's sample contains one anecdote from each route, $\gamma_\beta(\alpha)$, can be computed as

$$\gamma_\beta(\alpha) = \alpha\Big(\frac{\beta(1-\alpha)}{\alpha+\beta(1-\alpha)}\Big) + (1-\alpha)\Big(\frac{\beta\alpha}{\beta\alpha+(1-\alpha)}\Big). \tag{4}$$

Among customers who receive one anecdote from each route, let $\phi(\alpha)$ be the fraction who choose route SF, i.e, the fraction whose SF anecdote is smaller than their FS anecdote. We characterize $\phi(\alpha)$ explicitly in Appendix E (see Lemma 6). By equations (3) and (4), we can express the response function π_β by

$$\pi_\beta(\alpha) = \begin{cases} \gamma_\beta(\alpha)\phi(\alpha) + \dfrac{1-\gamma_\beta(\alpha)}{2} & \text{if } \alpha \in (0,1), \\ \dfrac{1}{2} & \text{if } \alpha \in \{0,1\}. \end{cases} \tag{5}$$

The function $\phi(\alpha)$ is continuous, which by equation (5) implies that $\pi_\beta(\alpha)$ is also continuous.

We first establish the size of the gap between the case with $\beta = 1$, when there is no sampling bias in drawing the anecdote, and the herding equilibrium. Observe from the left panel of Figure 1 that when $\beta = 1$ and $\mu_S = \mu_F$, the only equilibrium is $\alpha = 1/2$. This is extremely far from herding. The reason for the discrepancy is that when $\beta = 1$, if α is far from $1/2$, then customers are highly likely to receive both anecdotes from the same route. Therefore, if α moves away from the interior, then more customers must randomize, absent any information about one of the routes. This randomization pushes the SF fraction back towards $1/2$ and away from herding. For general service rates, there is exactly one equilibrium, which we fully characterize with the next proposition.

PROPOSITION 2 **(Unique Equilibrium for $\beta = 1$).** *Under the procedurally rational model with $K = 2$ and $\beta = 1$, the equilibrium is unique and given by*

$$\alpha_1 = \begin{cases} \dfrac{1}{\sqrt{2}} & \text{if} \quad \dfrac{\mu_S}{\mu_F} \leq \dfrac{1}{\sqrt{2}}, \\ \dfrac{\mu_S}{\mu_S + \mu_F}\left(1 + \sqrt{\dfrac{(\mu_F - \mu_S)(2\mu_S + \mu_F)}{2\mu_S\mu_F}}\right) & \text{otherwise.} \end{cases} \tag{6}$$

Proposition 2 shows that the equilibrium is decreasing in the service rate ratio μ_S/μ_F for $\beta = 1$. In addition, across all service rates, the unique equilibrium is at its maximum of $1/\sqrt{2} \approx .707$ when the service rate ratio is less than $1/\sqrt{2}$, and it decreases to a minimum of exactly $1/2$ when $\mu_S = \mu_F$. So, in all cases, the gap between the equilibrium and herding is sizeable when discernibility is low.

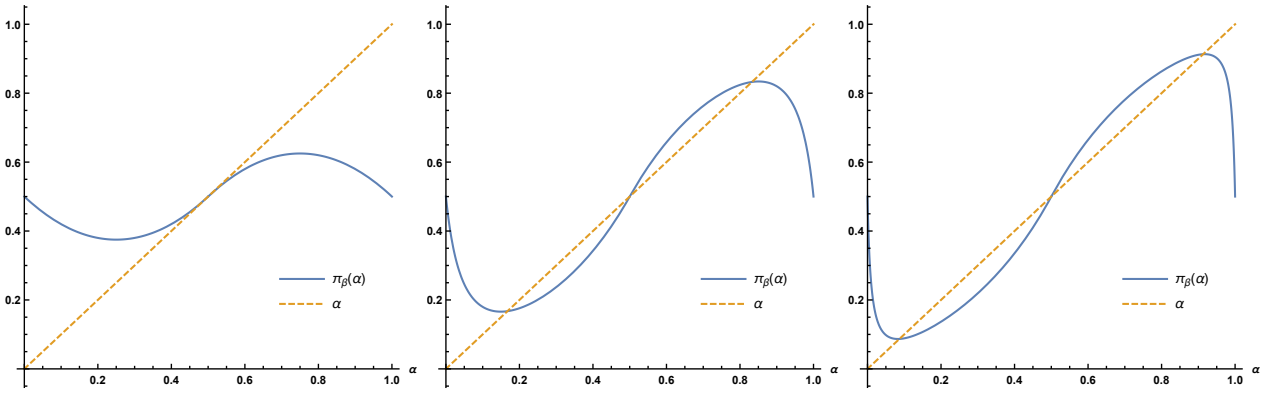
Figure 1 depicts the response function $\pi_\beta(\alpha)$ for different values of the discernibility parameter β with $\mu_S = \mu_F$. The far left panel of Figure 1 shows the "unbiased" case with $\beta = 1$. Naturally, for values of α close to 0 or 1, almost all customers receive both anecdotes from the same route, and these customers mix 50-50 between the routes. For larger values of β, customers are able to bias their second anecdote toward the missing route, so for a given α, a higher fraction of customers receive both types of anecdotes. The effect of this change is that the largest equilibrium α_β increases toward the fully rational outcome of herding (i.e., $\alpha = 1$) as β increases, as implied by Lemma 12 in Appendix E, which is a stronger version of Theorem 1 for the case of $K = 2$. Specifically, for the service rates in the figure ($\mu_S = \mu_F = 1$), we have $\alpha_1 = 1/2$, $\alpha_{20} = .833$, and $\alpha_{100} = .913$.

We will show in Section 5 that the cumulative system time of all customers is decreasing with α_β and that it is socially optimal for customers to herd. Hence, the equilibrium with low discernibility, which can be quite far from herding as shown by Proposition 2, can result in excess cumulative system time; in other words, the lower the discernibility, the worse the customer experience.

A caveat to our results about the largest equilibrium α_β is that there may exist other equilibria in a procedurally rational model. There are several reasons that the existence of multiple equilibria is not overly concerning. First, we believe that a service provider may be able to facilitate certain equilibrium outcomes, by methods such as providing incentives. Second, we have the following result showing that there is just one equilibrium in the interval $(1/2, 1]$.

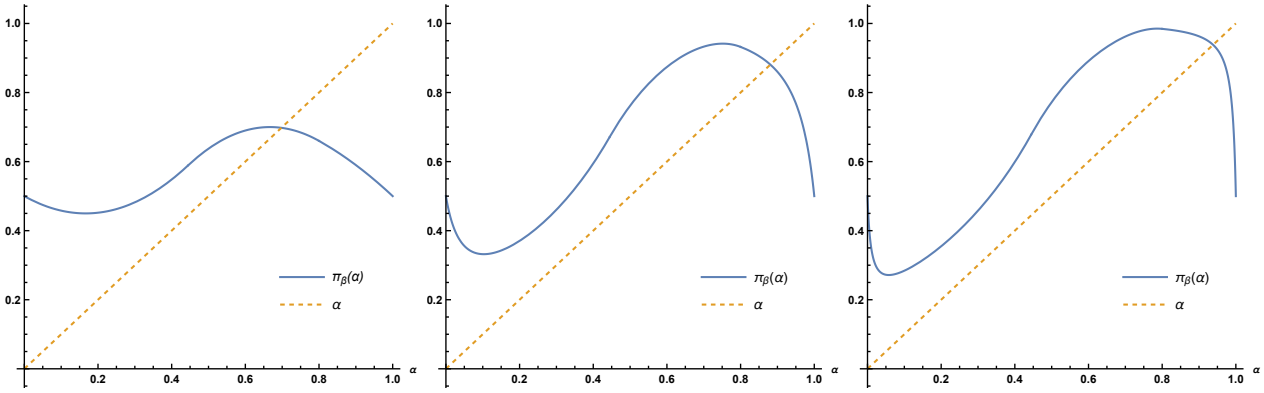
PROPOSITION 3 **(Exactly One Equilibrium Above 1/2).** *If $\beta \geq 1$ and $\mu_S < \mu_F$, or if $\beta > 1$ and $\mu_S = \mu_F$, then there is exactly one $\alpha \in (1/2, 1]$ satisfying $\pi_\beta(\alpha) = \alpha$.*

Third, for service rates that differ nonnegligibly, there is often a unique equilibrium. Indeed, for $K = 2$ and $\mu_S/\mu_F < c$ for $c \approx .93$, it can be proved that the equilibrium α_β is unique; we omit the



(a) $\beta = 1$ (b) $\beta = 20$ (c) $\beta = 100$

Figure 1 **Procedurally rational response function $\pi_\beta(\alpha)$ ($K = 2$, $\mu_S = 1$, $\mu_F = 1$).**



(a) $\beta = 1$ (b) $\beta = 20$ (c) $\beta = 100$

Figure 2 **Procedurally rational response function $\pi_\beta(\alpha)$ ($K = 2$, $\mu_S = .8$, $\mu_F = 1$).**

details for brevity, but the result follows straightforwardly using the explicit formula for $\phi(\alpha)$ found in Appendix E. See Figure 2 for examples. Qualitatively similar behavior also holds numerically for $K > 2$ (see Figure 7 in Section 6.2). These points further justify our focus on the largest equilibrium α_β. Next, we study the implications of discernibility for system performance.

5. Cumulative System Time

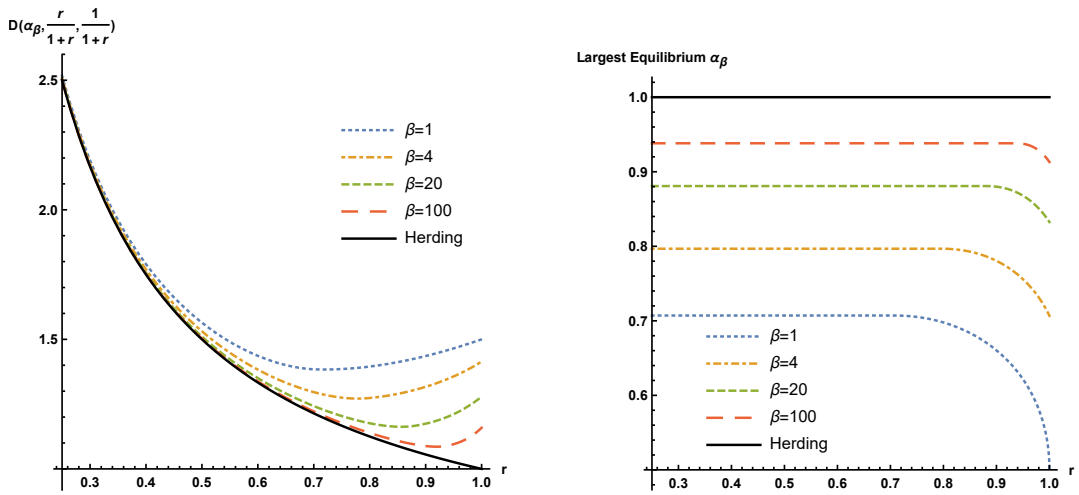
We now compare the performance of the system for different parameter values. We will be especially interested in the system times that customers experience at the largest equilibrium α_β, for different values of β and the service rates μ_S and μ_F. To measure how well the system performs—and by extension, how satisfied customers are likely to be—we use the *cumulative system time*, denoted by $D(\alpha, \mu_S, \mu_F)$. This quantity measures the cumulative delay by the integral of the total system time experienced by customers in each position in the queues. We provide a closed-form expression in Lemma 13 in Appendix F; importantly, the cumulative system time under herding is the same whether on route SF or route FS.

Arlotto et al. (2019) report that herding achieves excellent performance with respect to cumulative system time in the discrete setting. The next proposition verifies that in the fluid case also, herding performs extremely well and is, in fact, optimal.

PROPOSITION 4 **(Herding Is Socially Optimal)**. *The cumulative system time is minimized when customers herd, i.e., for fixed μ_S and μ_F, we have* $\arg\min_{\alpha\in[0,1]} D(\alpha,\mu_S,\mu_F)=\{0,1\}$.

Next, we look at how the relative cumulative system time (compared to the baseline cumulative time under herding with $\mu_S=\mu_F$) varies with different service rate ratios. Without loss of generality, we fix the sum of the service rates at 1. That is, we study the cumulative system time as a function of x, where $\mu_S=x$ and $\mu_F=1-x$. We can think of the service provider as having a certain amount of capacity (servers, machines, etc.) that he can divide among the two stations. We will compare the system performance in equilibrium between the fully rational case and the procedurally rational case as the allocation x changes, for different values of the discernibility β. It is convenient to relate x to the service rate ratio $r=\mu_S/\mu_F$. We have $r=x/(1-x)$, and therefore $x=r/(1+r)$. As x ranges from 0 to $1/2$, the ratio r ranges from 0 to 1.

Figure 3a illustrates how cumulative system time changes with service rate ratios under different values of β (like Section 4.2, we fix $K=2$, and note that we plot only the interval $[.25,1]$ because the system time increases without bound as the ratio approaches zero). It is clear that, if customers will herd (the solid line in Figure 3a), then the cumulative system time is minimized by setting the service rates exactly equal, i.e., $r=1$ and $\mu_S=\mu_F=1/2$. Interestingly, for procedurally rational customers, the cumulative system time is non-monotonic in the service rate ratio. The reason is that the largest equilibrium α_β is constant in an interval $[0,c]$ for some $c<1$ that varies with β. Below this threshold c, as we increase the service rate ratio, the system is more balanced in terms of service rates, but the equilibrium does not change (see Figure 3b), so there is unambiguous improvement in the cumulative system time. However, above this threshold, the procedurally rational equilibrium begins to decrease with the service rate ratio (again, see Figure 3b), which moves customers farther away from herding, so the cumulative system time curve reaches a minimum strictly below 1. By Proposition 4, moving away from herding will increase the cumulative system time for fixed service rates, and we observe that this effect outweighs the relative improvement from bringing the service rates closer together. Encouragingly, the attendant performance loss diminishes as discernibility improves. Now, suppose that we allow the firm to pick the optimal ratio r. From Figure 3a, for $\beta=1$, the cumulative system time at the optimal ratio is 38.5% worse than that at the optimal ratio under herding; by contrast, for $\beta=100$, the best-case cumulative system time is only 8.6% worse than the best case under herding. Put simply, in a service network with procedurally rational customers, the higher the discernibility, the better the customer experience.



(a) Cum. sys. time at largest equilibrium α_β (b) Largest equilibrium α_β

Figure 3 Cumulative system time and equilibria under herding and procedural rationality, versus $r = \mu_S/\mu_F$.

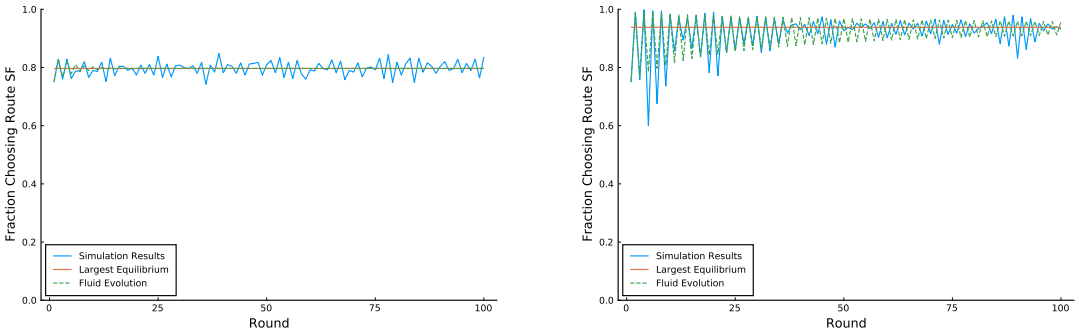
6. Numerical Studies

In this section, we conduct two numerical studies. In the first, we simulate a version of our model with atomic customers, which validates our findings from the fluid model. In the second, we numerically characterize the response function $\pi_{K,\beta}(\alpha)$ for larger sample sizes $K > 2$.

6.1. Systems with Atomic Customers

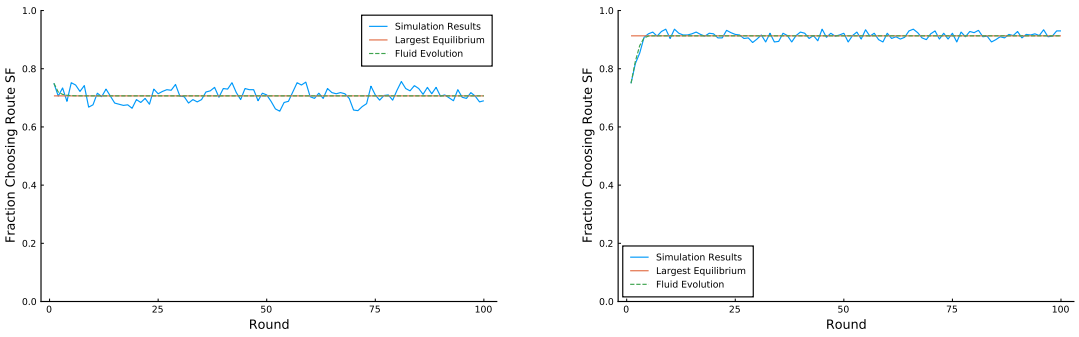
We now conduct a simulation study of systems with atomic customers under procedural rationality, to ascertain how closely the behavior of these systems resembles the fluid model. With atomic customers, we numerically study the evolution of procedurally rational routing decisions in a repeated setting. Specifically, we suppose that new customers enter to play the game in every period, and their anecdotes are drawn from customers in the previous period. As in Section 4.2, we assume that $K = 2$. If a fraction α takes route SF in a given period, then in the next period, a given customer's sample will contain anecdotes of both types with probability $\gamma_\beta(\alpha)$; with probability $1 - \gamma_\beta(\alpha)$, she receives two anecdotes of the same type and randomizes equally between the two routes. Because we are focusing on the largest equilibrium, and by Proposition 3 there is only one equilibrium for $\alpha \in (1/2, 1]$, we initialize $\alpha = .75$ in the first round, midway between $1/2$ and 1. For a range of service rates μ_S (we fix $\mu_F = 1$ in our simulations) and discernibility parameters β, we simulated systems with $N = 500$ customers for 100 rounds each, replicating each system 10 times.

First, we observe that, even for systems with only 500 customers, our analytical results for the fluid system are very good predictors of the routing decisions of procedurally rational customers. Table 1 reports sample statistics for the fraction of SF customers in rounds 75-100, across replications for each parameter combination. For reference, the value of the largest fluid equilibrium SF fraction is reported in the top right. We observe that play tends to hover very closely around the fluid



(a) $\beta = 4$ (b) $\beta = 100$

Figure 4 **Simulated evolution of discrete system ($N = 500, \mu_S = .6, \mu_F = 1$).**



(a) $\beta = 4$ (b) $\beta = 100$

Figure 5 **Simulated evolution of discrete system ($N = 500, \mu_S = 1, \mu_F = 1$).**

equilibrium. The medians are extremely close to α_β, and the first and third quartiles are between 0 and .08 above or below α. In addition, for given service rates, we observe that the interquartile range tends to be larger for larger values of β. The reason for this behavior is that the fluid evolution takes more rounds to converge as β increases.

Figures 4 and 5 show typical sample paths of the discrete system, for different values of the service rate μ_S and discernibility β. We observe a similar effect in these figures, namely that play quickly approaches the largest procedurally rational equilibrium α_β and then fluctuates in a relatively narrow band around it. Comparing the left and right panels of the figures, we can see that increasing β brings the largest equilibrium closer to herding. When β increases from 4 to 100, the largest equilibrium increases from .797 to .938 for $\mu_S = .6$, and from .707 to .913 for $\mu_S = 1$. In addition, for larger values of β, there is more fluctuation in the SF fraction. This is consistent with our observation from Table 1 that the interquartile range increases with β.

The reason for this behavior is that, as β increases and the equilibrium converges to herding, the equilibrium may be very close to the region where many customers fail to receive anecdotes of both routes, and the response π_β decreases steeply in this region. If the SF fraction enters this

Table 1 Fraction of SF customers in rounds 75-100 ($\mu_F = 1$, $N = 500$, 100 rounds, 10 replications)

	μ_S					
β	0.5	0.6	0.7	0.8	0.9	1
1	0.708	0.706	0.707	0.696	0.656	0.528
4	0.796	0.794	0.796	0.796	0.778	0.708
20	0.878	0.881	0.880	0.882	0.880	0.830
100	0.937	0.939	0.940	0.938	0.938	0.912

(a) Median

	μ_S					
β	0.5	0.6	0.7	0.8	0.9	1
1	0.707	0.707	0.707	0.698	0.661	0.500
4	0.797	0.797	0.797	0.797	0.781	0.707
20	0.881	0.881	0.881	0.881	0.880	0.833
100	0.938	0.938	0.938	0.938	0.938	0.913

(b) Largest Fluid Model Equilibrium (α_β)

	μ_S					
β	0.5	0.6	0.7	0.8	0.9	1
1	0.688	0.690	0.690	0.684	0.642	0.500
4	0.774	0.772	0.774	0.778	0.768	0.694
20	0.818	0.852	0.851	0.840	0.864	0.820
100	0.857	0.884	0.846	0.906	0.916	0.904

(c) First Quartile

	μ_S					
β	0.5	0.6	0.7	0.8	0.9	1
1	0.726	0.722	0.722	0.708	0.674	0.552
4	0.816	0.819	0.820	0.811	0.792	0.720
20	0.925	0.902	0.908	0.912	0.892	0.842
100	0.976	0.968	0.980	0.960	0.956	0.922

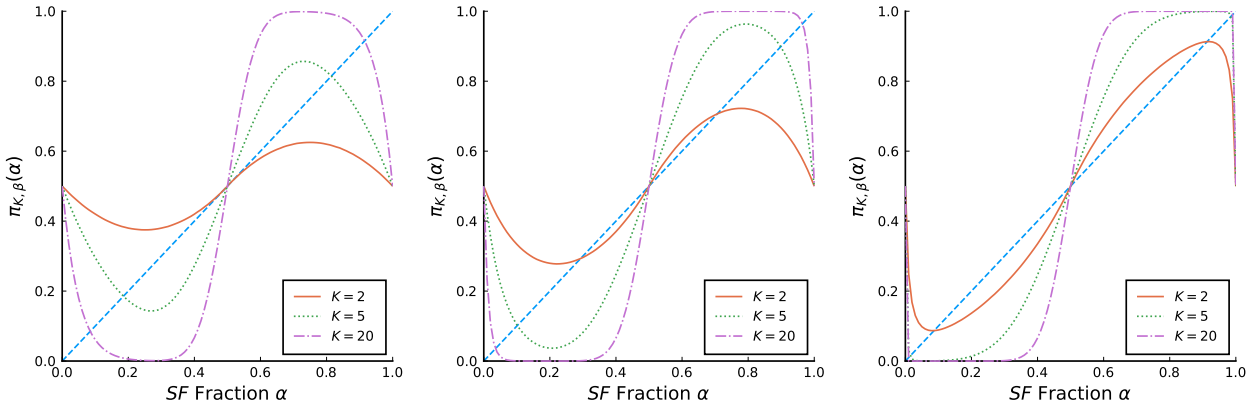
(d) Third Quartile

steeply decreasing region above the equilibrium, then the larger number of randomizing customers can cause a reflection in the next period that pushes the SF fraction back below the equilibrium. Accordingly, for large β, we observe an oscillating SF fraction around the largest equilibrium in the fluid evolution, with the oscillations decreasing in amplitude as play converges toward the equilibrium. The discrete system exhibits similar behavior, with some additional randomness in the timing of the oscillations and their amplitudes. This effect attenuates for service rates that are closer together, as seen in the right panel of Figure 5: with equal service rates, even though the equilibrium is close to herding, there is less oscillation. This observation accords with Table 1: for given β, this range becomes smaller for service rate ratios closer to 1. Overall, these results reflect similar behavior to that predicted by the fluid approximation. We next consider the case of larger sample sizes and numerically characterize the response function $\pi_{K,\beta}(\alpha)$.

6.2. Larger Sample Sizes

As the sample size grows, exact analytical characterization of the response function $\pi_{K,\beta}(\alpha)$ quickly becomes intractable. To augment our structural results, we now numerically compute the response function for various values of the sample size K and the discernibility parameter β.

We compute the response function via Monte Carlo simulation: for given K and β and on a grid of α between 0 and 1, we simulate our sampling process repeatedly and record the fraction of such samples in which the SF sample average system time is less than the FS sample average system time; the result is our computed value for $\pi_{K,\beta}(\alpha)$. Each panel of Figures 6 and 7 plots $\pi_{K,\beta}(\alpha)$ for three different sample sizes, and different panels reflect different values of the discernibility parameter β. Figure 6 is for equal service rates at both stations, and Figure 7 depicts unequal



(a) $\beta = 1$ (b) $\beta = 4$ (c) $\beta = 100$

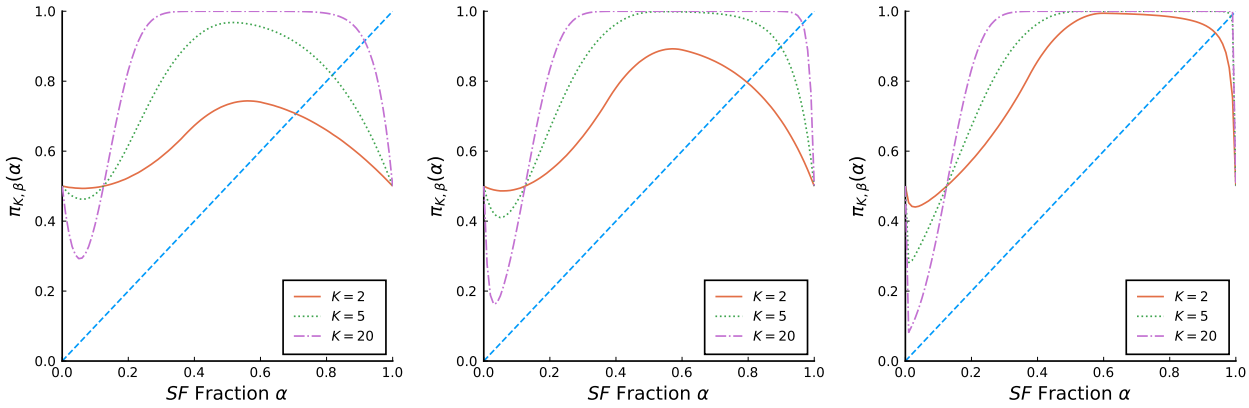
Figure 6 **Procedurally rational response function $\pi_{K,\beta}(\alpha)$ ($K = 2$, $\mu_S = 1$, $\mu_F = 1$).**

service rates. The line $f(\alpha) = \alpha$ is also plotted for visual reference, and intersections with this line correspond to procedurally rational equilibria for the given parameters.

In the leftmost panel of Figure 6 we have $\beta = 1$, and as expected, the largest equilibrium $\alpha_{K,\beta}$ is far from herding for small K. However, as K increases, $\alpha_{K,\beta}$ also increases. In fact, for values of α that are close but not too close to 1, the response function with $K = 20$ is visually indistinguishable from herding. But as α approaches 1, even with $K = 20$ it becomes more probable that all anecdotes will come from route SF, which causes the response function to reduce towards 1/2. For this reason, the equilibrium with $\beta = 1$ and $K = 20$ is relatively close to herding, but is still non-negligibly different from it, even though we have $\pi_{20,1}(\alpha) \approx 1$ for some values of α. Nonetheless, that $\alpha_{K,\beta}$ increases toward herding as K increases aligns with Proposition 1, and increasing the sample size has a similar effect for larger discernibility in the other two panels. Now, comparing panels from left to right, we see that for each sample size K, as the discernibility β increases, the largest equilibrium $\alpha_{K,\beta}$ also increases. Additionally, recall from Section 4.2 that the equilibrium is unique for $K = 2$ when the service rates are not too close together. In Figure 7, we observe a similar phenomenon (i.e., differing service rates yielding a unique equilibrium) for larger sample sizes. In all cases and in both figures, the comparative statics with more than two anecdotes are similar to those for $K = 2$. Most importantly, and consistent with our earlier results, we find for each sample size and in both figures that as discernibility increases, the largest equilibrium inevitably approaches herding.

7. Conclusion

We study an open-routing service network with two stations and self-interested, procedurally rational customers who make decisions about which route to take through the network based on anecdotes. In the fully rational counterpart to our model, customers herd, i.e., they take the same route through



(a) $\beta = 1$ (b) $\beta = 4$ (c) $\beta = 100$

Figure 7 **Procedurally rational response function $\pi_{K,\beta}(\alpha)$ ($K = 2$, $\mu_S = 0.6$, $\mu_F = 1$).**

the network; and this outcome is also socially optimal in terms of cumulative system time. We find that, if procedurally rational customers are sufficiently likely to receive anecdotes about both routes (corresponding to high discernibility in our model), then their decisions will approach herding, i.e., the social optimum. By contrast, if the types of their anecdotes are drawn proportional to the population (low discernibility), then the equilibrium strays far from herding, resulting in much longer waits for customers. A key takeaway from our work for managers of open-routing service networks is that, when customers can choose from multiple routes, their experience is dictated not so much by the sample *size* but instead by how likely they are to obtain anecdotes about all routes. Surprisingly, even if each customer's sample is quite small, limited but diverse information encompassing all of the alternatives can still significantly impact customer decisions towards the socially optimal outcome of herding. Managers of open-routing service networks should thus be keenly aware of the importance of discernibility, and if they can facilitate information sharing and more active online communities, e.g., by creating and promoting online platforms for customers to share their experiences, then it is advisable to do so.

We believe that the value of information diversity, even in the face of limited information quantity, may have broader applicability beyond the setting of this work. Generally speaking, customers who must make decisions based on limited information about their alternatives are fundamentally crippled if their information does not extend to one or more of their possible actions, and service providers should account for this limitation on customers when planning for their decisions. We remark, however, that the fully rational outcome may not always be the goal for managers. There are many examples in the literature in which selfish customer behavior leads to socially sub-optimal outcomes, and the cost of this selfish behavior has even been formalized in the notion of a price of anarchy (see, e.g., Roughgarden and Tardos 2002). Indeed, in settings where the fully rational

outcome is undesirable, if customers exhibit procedural rationality, then to improve social welfare, managers might actually wish to *decrease* discernibility, making it *less* likely that a customer's sample will include anecdotes about all of the alternatives. Accordingly, the new sequential sampling procedure that we introduce could be applied to other contexts in which customers exhibit procedural rationality, possibly to assess when high or low discernibility is better for social welfare. Because the procedure extends the procedural rationality framework by incorporating discernibility, we believe that it is also of independent interest to researchers studying bounded rationality, particularly in a service setting.

Beyond the broad strokes provided in the previous paragraph, we identify several opportunities for further research into open-routing service systems with procedurally rational customers. First, we believe that a promising direction would be to take our findings to the laboratory. It would be of interest to test whether subjects exhibit procedural rationality when faced with navigating an open-routing service network. Also valuable would be to provide anecdotes to subjects according to our sequential sampling procedure and measure their responses. Finally, while our model considers customers who are all present at the start of service, it could be worthwhile to investigate procedural rationality in a network similar to ours but to which customers arrive over time. Such a study could present an additional interesting challenge in modelling the temporal component of the sampling procedure. Given that such a network would evolve over time according to recent arrivals and decisions, it is not obvious from what interval of time would customers' samples be drawn, and whether their decision process would incorporate the "recentness" of the sample. To analytically study these or other complex systems (e.g., systems with more than two stations) would be extremely challenging, whether with fully rational or procedurally rational customers. However, such research could reveal valuable insights, and we also note that our framework at least provides a basis for numerical study of more complex systems. Overall, we hope that our findings will encourage others to pursue similar research, whether on open routing specifically or, more broadly, on procedural rationality and discernibility in service contexts.

References

Acemoglu D, Makhdoumi A, Malekian A, Ozdaglar A (2018) Informational Braess' Paradox: The effect of information on traffic congestion. *Operations Research* 66(4):893–917.

Arlotto A, Frazelle AE, Wei Y (2019) Strategic open routing in service networks. *Management Science* 65(2):735–750.

Baron O, Berman O, Krass D, Wang J (2016) Strategic idleness and dynamic scheduling in an open-shop service network: Case study and analysis. *Manufacturing & Service Operations Management* 19(1):52–71.

Boyd S, Vandenberghe L (2004) *Convex Optimization* (Cambridge University Press).

Braess D (1968) Über ein paradoxon aus der verkehrsplanung. *Unternehmensforschung* 12(1):258–268.

Çil EB, Lariviere MA (2013) Saving seats for strategic customers. *Operations Research* 61(6):1321–1332.

Chen Y, Su X, Zhao X (2012) Modeling bounded rationality in capacity allocation games with the quantal response equilibrium. *Management Science* 58(10):1952–1962.

Cohen JE, Kelly FP (1990) A paradox of congestion in a queuing network. *Journal of Applied Probability* 730–734.

Cui S, Su X, Veeraraghavan S (2019) A model of rational retrials in queues. *Operations Research* 67(6):1699–1718.

Dai T, Zheng R, Sycara K (2020) Jumping the line, charitably: Analysis and remedy of donor-priority rule. *Management Science* 66(2):622–641.

Debo L, Snitkovsky RI (2018) Tipping in service systems: The role of a social norm. *Working Paper, Dartmouth College* .

Donkers B, Franses PH, Verhoef PC (2003) Selective sampling for binary choice models. *Journal of Marketing Research* 40(4):492–497.

Fiedler K (2000) Beware of samples! A cognitive-ecological sampling approach to judgment biases. *Psychological review* 107(4):659.

Fiedler K, Juslin P, et al. (2006) *Information Sampling and Adaptive Cognition* (Cambridge University Press).

Grimmett GS, Stirzaker D (2020) *Probability and Random Processes* (Oxford University Press).

Hassin R (2016) *Rational Queueing* (Chapman and Hall/CRC).

Hassin R, Haviv M (2003) *To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems* (Springer Science & Business Media).

He H, Garcia EA (2009) Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering* 21(9):1263–1284.

Honnappa H, Jain R (2015) Strategic arrivals into queueing networks: The network concert queueing game. *Operations Research* 63(1):247–259.

Huang T, Allon G, Bassamboo A (2013) Bounded rationality in service systems. *Manufacturing & Service Operations Management* 15(2):263–279.

Huang T, Chen YJ (2015) Service systems with experience-based anecdotal reasoning customers. *Production and Operations Management* 24(5):778–790.

Huang T, Yin Z, Chen YJ (2017) Managing posterior price matching: The role of customer boundedly rational expectations. *Manufacturing & Service Operations Management* 19(3):385–402.

Huang T, Yu Y (2014) Sell probabilistic goods? A behavioral explanation for opaque selling. *Marketing Science* 33(5):743–759.

Juslin P, Winman A, Hansson P (2007) The naïve intuitive statistician: A naïve sampling model of intuitive confidence intervals. *Psychological review* 114(3):678.

Kremer M, Debo L (2016) Inferring quality from wait time. *Management Science* 62(10):3023–3038.

Li X, Guo P, Lian Z (2016) Quality-speed competition in customer-intensive services with boundedly rational customers. *Production and Operations Management* 25(11):1885–1901.

Liu Y, Wang X, Gilbert S, Lai G (2018) Pricing, quality and competition at on-demand healthcare service platforms. *Working Paper, University of Texas, Austin* .

Nageswaran L, Scheller-Wolf A (2022) Queues with redundancy: Is waiting in multiple lines fair? *Manufacturing & Service Operations Management* 24(4):1959–1976.

Naor P (1969) The regulation of queue size by levying tolls. *Econometrica: Journal of the Econometric Society* 15–24.

Oh J, Su X (2018) Reservation policies in queues: Advance deposits, spot prices, and capacity allocation. *Production and Operations Management* 27(4):680–695.

Osborne MJ, Rubinstein A (1998) Games with procedurally rational players. *American Economic Review* 834–847.

Owen A, Zhou Y (2000) Safe and effective importance sampling. *Journal of the American Statistical Association* 95(449):135–143.

Parlaktürk AK, Kumar S (2004) Self-interested routing in queueing networks. *Management Science* 50(7):949–966.

Ren H, Huang T (2018) Modeling customer bounded rationality in operations management: A review and research opportunities. *Computers & Operations Research* 91:48–58.

Ren H, Huang T, Arifoglu K (2018) Managing service systems with unknown quality and customer anecdotal reasoning. *Production and Operations Management* 27(6):1038–1051.

Roughgarden T, Tardos É (2002) How bad is selfish routing? *Journal of the ACM (JACM)* 49(2):236–259.

Shtrichman O, Ben-Haim R, Pollatschek MA (2001) Using simulation to increase efficiency in an army recruitment office. *Interfaces* 31(4):61–70.

Smith L, Sørensen PN (2020) Rational social learning with random sampling. *Working Paper, University of Wisconsin* .

Spectrio (2015) *Queue Management and the Customer Experience.* `https://blogs.spectrio.com/queue-management-and-the-customer-experience`, Accessed 10/8/2019.

Spiegler R (2006) The market for quacks. *The Review of Economic Studies* 73(4):1113–1131.

Su X (2008) Bounded rationality in newsvendor models. *Manufacturing & Service Operations Management* 10(4):566–589.

Su X, Zenios S (2004) Patient choice in kidney allocation: The role of the queueing discipline. *Manufacturing & Service Operations Management* 6(4):280–301.

Tong J, Feiler D (2017) A behavioral model of forecasting: Naive statistics on mental samples. *Management Science* 63(11):3609–3627.

Tversky A, Kahneman D (1971) Belief in the law of small numbers. *Psychological Bulletin* 76(2):105–110.

Veeraraghavan SK, Debo LG (2011) Herding in queues with waiting costs: Rationality and regret. *Manufacturing & Service Operations Management* 13(3):329–346.

Wu S, Liu Q, Zhang RQ (2015) The reference effects on a retailer's dynamic pricing and inventory strategies with strategic consumers. *Operations Research* 63(6):1320–1335.

Yang L, Debo L (2019) Referral priority program: Leveraging social ties via operational incentives. *Management Science* 65(5):2231–2248.

Yang L, Guo P, Wang Y (2018) Service pricing with loss-averse customers. *Operations Research* 66(3):761–777.

Online Appendix. Discussion, Auxiliary Results, Proofs, and Extension

This appendix is divided into sections. First, we provide interpretation of the discernibility parameter β in Appendix A. Next, we provide technical details of our fluid approximation in Appendix B. In Appendix C, we prove Claim 1. After that, we provide the proofs of the major results (propositions or theorems) from Sections 4.1, 4.2, and 5 of the main body in Appendices D, E, and F, respectively, including any auxiliary results or corollaries. Finally, in Appendix G, we show that our main finding continues to hold if we relax the assumption of equal-probability randomization under missing anecdotes.

A. Interpretation for the Discernibility Parameter

First, we point out that sampling probabilities with discernibility are similar to the logit choice model, where β can be viewed as the parameter that changes the odds of drawing SF or FS from uniform selection.

Second, we can interpret the probabilities from (1) and (2) as the probability of drawing FS (or SF) type, when the customer is drawing customers with feature X, where the feature is biased towards FS (or SF). Imagine that a customer, after observing that her first anecdote is from route SF, draws the next anecdote by selecting customers with feature X. Then the probability that she draws her next sample from route FS, by Bayes Theorem, is

$$\begin{aligned}\mathsf{Pr}(FS|X) = \frac{\mathsf{Pr}(FS)\mathsf{Pr}(X|FS)}{\mathsf{Pr}(X)} &= \frac{(1-\alpha)\mathsf{Pr}(X|FS)}{\mathsf{Pr}(X,FS)+\mathsf{Pr}(X,SF)} \\ &= \frac{(1-\alpha)\mathsf{Pr}(X|FS)}{(1-\alpha)\mathsf{Pr}(X|FS)+\alpha\mathsf{Pr}(X|SF)} \\ &= \frac{(1-\alpha)\mathsf{Pr}(X|FS)/\mathsf{Pr}(X|SF)}{(1-\alpha)\mathsf{Pr}(X|FS)/\mathsf{Pr}(X|SF)+\alpha}.\end{aligned}$$

Finally, if we let $\beta = \mathsf{Pr}(X|FS)/\mathsf{Pr}(X|SF)$, then

$$\mathsf{Pr}(FS|X) = \frac{(1-\alpha)\beta}{\alpha+(1-\alpha)\beta}.$$

B. Fluid Approximation Details

Given its service rate μ_S, in an elapsed time of length ℓ, station S is capable of processing a volume $\mu_S\ell$ of fluid. Correspondingly, to process a volume v of fluid at station S requires a length of time equal to

$$T_S(v) = \frac{v}{\mu_S}. \tag{7}$$

We similarly have that in an elapsed time of length ℓ, station F can process a volume $\mu_F\ell$ of fluid. To processes a volume v of fluid at station S requires a length of time equal to

$$T_F(v) = \frac{v}{\mu_F}. \tag{8}$$

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.

Recall that we use α to denote the fraction α of customers that chooses route SF. Denote by $Q_S(\alpha;\ell)$ the amount of fluid waiting in the queue (or "buffer") at station S given an SF fraction of α, an elapsed time

ℓ after the system begins operating. Define $Q_F(\alpha;\ell)$ similarly for station F. Thus, we have $Q_S(\alpha;0)=\alpha$, and $Q_F(\alpha;0)=1-\alpha$. Letting $[x]^+=\max\{x,0\}$, the amount of fluid $Q_F(\alpha;\ell)$ in station F's buffer after the system has been operating for an elapsed time ℓ is equal to

$$\begin{aligned} Q_F(\alpha;\ell) &:= \Big[Q_F(\alpha;0)+\min\{Q_S(\alpha;0),\mu_S\ell\}-\mu_F\ell\Big]^+ \\ &= \Big[1-\alpha+\min\{\alpha,\mu_S\ell\}-\mu_F\ell\Big]^+. \end{aligned} \tag{9}$$

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

$$\begin{aligned} Q_S(\alpha;\ell) &:= \Big[Q_S(\alpha;0)+\min\{Q_F(\alpha;0),\mu_F\ell\}-\mu_S\ell\Big]^+ \\ &= \Big[\alpha+\min\{1-\alpha,\mu_F\ell\}-\mu_S\ell\Big]^+. \end{aligned} \tag{10}$$

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

Let y_S be a possible starting position in the buffer at station S, where $0\le y_S\le\alpha$, and let y_F be a possible starting position in the queue at station F, where $0\le y_F\le 1-\alpha$. We denote by $\mathcal{S}(\alpha;y_S)$ the total system time for the infinitesimal customer starting in position y_S in the station S queue, given a total SF fraction α.

LEMMA 1 **(System Time for *SF* Customers).** *The function $\mathcal{S}(\alpha;y_S)$ can be expressed as*

$$\mathcal{S}(\alpha;y_S)=\begin{cases} \dfrac{y_S+1-\alpha}{\mu_F} & \text{if } y_S\le\mu_S\Big(\dfrac{1-\alpha}{\mu_F-\mu_S}\Big), \\ \dfrac{y_S}{\mu_S} & \text{otherwise.} \end{cases} \tag{11}$$

When $\mu_S=\mu_F$, we treat $\frac{1-\alpha}{\mu_F-\mu_S}$ as positive infinity, implying that $\mathcal{S}(\alpha;y_S)=\frac{y_S+1-\alpha}{\mu_F}$.

Proof. The fluid at position y_S in the buffer at station S will depart from station S after $T_S(y_S)=y_S/\mu_S$ units of time by equation (7). When this customer arrives at station F, the amount of fluid in the buffer there is $Q_F(\alpha;y_S/\mu_S)$, which we can determine using equation (9). We can therefore write her total system time $\mathcal{S}(\alpha;y_S)$ as

$$\begin{aligned} \mathcal{S}(\alpha;y_S) &= \frac{y_S}{\mu_S}+T_F\Big(Q_F\Big(\alpha;\frac{y_S}{\mu_S}\Big)\Big) \\ &= \frac{y_S}{\mu_S}+\frac{1}{\mu_F}\Big[1-\alpha+\min\{\alpha,y_S\}-\frac{\mu_F y_S}{\mu_S}\Big]^+ \\ &= \frac{y_S}{\mu_S}+\frac{1}{\mu_F}\Big[1-\alpha+y_S-\frac{\mu_F y_S}{\mu_S}\Big]^+, \end{aligned} \tag{12}$$

where the last substitution follows from the fact that for any SF customer we must have $y_S\le\alpha$. The bracketed term in equation (12) is nonnegative only if $y_S\le\mu_S(1-\alpha)/(\mu_F-\mu_S)$, in which case the inequality holds for any $\hat{y}_S$). Note that we define the RHS of the inequality as ∞ if $\mu_S=\mu_F$, for notational convenience: the inequality always holds in that case, correctly implying that the first piece of equation (7) always governs the function $\mathcal{S}$ for equal service rates. Substitution then yields the first expression in equation (11). Otherwise, i.e., if $y_S>\mu_S(1-\alpha)/(\mu_F-\mu_S)$, taking the positive part of the bracketed term gives zero, and we are left with y_S/μ_S. We conclude equation (11). □

The piecewise nature of $\mathcal{S}(\alpha; y_S)$ arises because the buffer at station F empties faster than that at station S. A customer near the front of the station S buffer will depart station S to find some customers still in the station F buffer, so her system time is determined by how long it takes station F to process both all of the FS customers and the SF customers in front of her. By contrast, a customer near the back of the station S buffer will arrive to station F after it clears and find it empty; for this infinitesimal customer, since her own service times are negligible, her system time is determined by how long station S takes to process the SF customers in front of her.

We similarly denote by $\mathcal{F}(\alpha; y_F)$ the total system time for the infinitesimal customer starting in position y_F in the station F queue, given a total SF fraction α. Because the station S buffer empties at a slower rate, all FS customers will be delayed in this buffer when they depart station F. The system time for these customers can thus be expressed more simply.

LEMMA 2 **(System Time for FS Customers)**. *We can express the function $\mathcal{F}(\alpha; y_F)$ by*

$$\mathcal{F}(\alpha; y_F) = \frac{\alpha + y_F}{\mu_S}. \tag{13}$$

Proof. The proof proceeds along similar lines to the proof of Lemma 1. The fluid at position y_F in the buffer at station F will depart from station F after $T_F(y_F) = y_F/\mu_F$ units of time by equation (8). When this customer arrives at station S, the amount of fluid in the buffer there is $Q_S(\alpha; y_F/\mu_F)$, which we can determine using equation (10). The total system time for this customer is then

$$\begin{aligned}
\mathcal{F}(\alpha; y_F) &= \frac{y_F}{\mu_F} + T_S\Big(Q_S\Big(\alpha; \frac{y_F}{\mu_F}\Big)\Big) \\
&= \frac{y_F}{\mu_F} + \frac{1}{\mu_S}\Big[\alpha + \min\{1-\alpha, y_F\} - \frac{\mu_S y_F}{\mu_F}\Big]^+ \\
&= \frac{y_F}{\mu_F} + \frac{1}{\mu_S}\Big(\alpha + y_F - \frac{\mu_S y_F}{\mu_F}\Big) \\
&= \frac{\alpha + y_F}{\mu_S}.
\end{aligned}$$

The equivalence between the second and third equations holds because we must have $y_F \le 1-\alpha$, and because $\mu_S/\mu_F \le 1$ then implies that the bracketed term is nonnegative. □

C. Proof of Claim 1

Proof. Suppose $\alpha = 1$, and consider an arbitrary customer contemplating deviation from route SF to route FS. For $\alpha = 1$ and $\mu_S < \mu_F$, the RHS of the condition for the first piece of equation (11) is zero, so we can focus on the second piece. Note that if $\mu_S = \mu_F$, although the first piece always governs the function, the expressions for the first and second piece are equal when $\alpha = 1$. Thus, we can use the second piece of equation (11) for our analysis for all $\mu_S \le \mu_F$. For a given starting position y_S in the station S buffer, the system time is then y_S/μ_S by Lemma 1. Because customers in each buffer are sequenced uniformly at random, the expected system time for an SF customer given $\alpha = 1$ is equal to $1/(2\mu_S)$. If an arbitrary nonatomic customer deviates to route FS, then her position in the station F buffer will be $y_F = 0$. By Lemma 2, her expected system time will thus be $1/\mu_S > 1/(2\mu_S)$, implying that she has no incentive to deviate to route FS. We conclude that $\pi^*(1) = 1$.

Next, suppose $\alpha = 0$ and $2\mu_S \ge \mu_F$. First, we note that if $\mu_S = \mu_F$, then $\pi^*(0) = 0$ by symmetry. For the rest of the proof, we assume that $\mu_S < \mu_F$. If a customer contemplating deviation remains on route FS, then by Lemma 2 and the fact that her starting position is uniform in the station F buffer, her expected system time is $1/(2\mu_S)$. If she deviates to route SF, then her position in the station S buffer will be $y_S = 0$. By Lemma 1, then, her system time will be $1/\mu_F \ge 1/(2\mu_S)$, where the inequality holds by our assumption that $2\mu_S \ge \mu_F$. Thus, the customer has no incentive to deviate from route FS, and we conclude that $\pi^*(0) = 0$. □

D. Proofs for Section 4.1

The proof of Proposition 1 requires three auxiliary lemmas. We first show with Lemma 3 that the response function $\pi_{K,\beta}$ is continuous in α. We then prove with Lemma 4 that the set of procedurally rational equilibria is non-empty and compact. Most of what remains to prove Proposition 1 is then accomplished by Lemma 5, which shows that for $\alpha \in (1/2, 1)$ and fixed β, as K increases, eventually both (i) customers will receive many anecdotes from both routes and (ii) their corresponding estimates of the system times for each route will be accurate enough that they will choose route SF with high probability. From there, the proof of Proposition 1 is straightforward. Following the proof of the proposition, we prove Corollary 1, that for any $\alpha \in (1/2, 1)$ and fixed β, the procedurally rational response function $\pi_{K,\beta}$ converges to the fully rational response function π^* (which prescribes herding, i.e., $\alpha = 1$) as $K \to \infty$. Lastly, we give the proof for Theorem 1.

LEMMA 3 **(Continuity of $\pi_{K,\beta}$).** *For any fixed K, β, the function $\pi_{K,\beta}(\alpha)$ is continuous over $[0, 1]$.*

Proof. For simplicity of notation, we drop K and β from π and γ as they are fixed. First, because the probability of each draw being of type SF (conditioned on the current number of SF-type anecdotes) is continuous in α, $\gamma(\alpha)$ is continuous in α in $[0, 1]$. By equations (1) and (2), for $\alpha = 0$ or 1 and any value of β, the probability of the unchosen route being included in the sample is zero, i.e., we have $\gamma(0) = \gamma(1) = 0$. As $\mathsf{Pr}[\bar{s}_{K,\beta}(\alpha) \leq \bar{f}_{K,\beta}(\alpha) | \mathcal{B}]$ is bounded within $[0, 1]$, we have

$$\lim_{\alpha \to 0^+} \pi(\alpha) = \lim_{\alpha \to 1^-} \pi(\alpha) = \frac{1}{2} = \pi(0) = \pi(1). \tag{14}$$

Therefore, $\pi(\alpha)$ is continuous at 0 and 1.

Next, we show that $\pi(\alpha)$ is continuous within $(0, 1)$. Let $X(\alpha)$ be the (random) number of SF anecdotes drawn from a sample, and $p_n(\alpha)$ be the probability that $X(\alpha) = n$. Then, we have that

$$\pi(\alpha) = \sum_{n=1}^{K-1} p_n(\alpha) \mathsf{Pr}[\bar{s}_{K|n}(\alpha) \leq \bar{f}_{K|n}(\alpha)] + \frac{1}{2}(1 - \gamma(\alpha)),$$

where $\bar{s}_{K|n}(\alpha)$ and $\bar{f}_{K|n}(\alpha)$ denote the conditional average system time for the SF (FS) anecdotes, given an SF fraction α and *conditioned on drawing exactly n SF anecdotes* (and thus exactly $K - n$ FS anecdotes). We prove that $\pi(\alpha)$ is continuous over $(0, 1)$ by proving that $\mathsf{Pr}[\bar{s}_{K|n}(\alpha) \leq \bar{f}_{K|n}(\alpha)]$ is continuous in α.

Observe that for each fixed n, $\bar{s}_{K|n}(\alpha)$ and $\bar{f}_{K|n}(\alpha)$ are independent sums of n and $K - n$ i.i.d. continuous random variables, respectively. Moreover, the probability density function of the system time for a randomly drawn SF anecdote is "continuous" with respect to α in $(0, 1)$ in the following sense: let $\hat{s}(\alpha)$ (and $\hat{s}(\alpha + \epsilon)$) be the system time for a randomly drawn SF anecdote given that α (and $\alpha + \epsilon$) chose SF. Then for any $-1 < \epsilon < 1$ such that $\alpha + \epsilon \in (0, 1)$, the shared area under the probability density functions of $\hat{s}(\alpha + \epsilon)$ and $\hat{s}(\alpha)$ is at least $1 - g(\epsilon)$, for some function g such that $\lim_{\epsilon \to 0} g(\epsilon) = 0$. Therefore, we can find a coupling through a map $\mathcal{T}_S$ such that $\mathcal{T}_S(\hat{s}(\alpha))$ has the same distribution as $\hat{s}(\alpha + \epsilon)$, and

$$\mathcal{T}_S(\hat{s}(\alpha)) \stackrel{d}{=} \hat{s}(\alpha + \epsilon),\ \mathsf{Pr}[\mathcal{T}_S(\hat{s}(\alpha)) = \hat{s}(\alpha)] \geq 1 - g(\epsilon). \tag{15}$$

Similarly, for any $-1 < \epsilon < 1$ such that $\alpha + \epsilon \in (0, 1)$, let $\hat{f}(\alpha)$ (and $\hat{f}(\alpha + \epsilon)$) be the system time for a randomly drawn FS anecdote given that a fraction α (and $\alpha + \epsilon$) chose SF. We can also find a coupling through a map $\mathcal{T}_F$ such that $\mathcal{T}_F(\hat{f}(\alpha))$ has the same distribution as $\hat{f}(\alpha + \epsilon)$, and

$$\mathcal{T}_F(\hat{f}(\alpha)) \stackrel{d}{=} \hat{f}(\alpha + \epsilon),\ \mathsf{Pr}[\mathcal{T}_F(\hat{f}(\alpha)) = \hat{f}(\alpha)] \geq 1 - g(\epsilon). \tag{16}$$

By equations (15) and (16), we get that there exists a coupling process through $\mathcal{T}_S$ such that

$$\mathsf{Pr}[\bar{s}'_{K|n}(\alpha) = \bar{s}'_{K|n}(\alpha+\epsilon)] \geq (1-g(\epsilon))^n,$$

where $\bar{s}'_{K|n}(\alpha)$ and $\bar{s}'_{K|n}(\alpha+\epsilon)$ have the same distributions as $\bar{s}_{K|n}(\alpha)$ and $\bar{s}_{K|n}(\alpha+\epsilon)$, respectively. Similarly, there exists another coupling process (independent of the first) through $\mathcal{T}_F$ such that

$$\mathsf{Pr}[\bar{f}'_{K|n}(\alpha) = \bar{f}'_{K|n}(\alpha+\epsilon)] \geq (1-g(\epsilon))^{K-n},$$

where $\bar{f}'_{K|n}(\alpha)$ and $\bar{f}'_{K|n}(\alpha+\epsilon)$ have the same distributions as $\bar{f}_{K|n}(\alpha)$ and $\bar{f}_{K|n}(\alpha+\epsilon)$, respectively. This in turn implies that

$$\mathsf{Pr}[\bar{s}'_{K|n}(\alpha) = \bar{s}'_{K|n}(\alpha+\epsilon), \bar{f}'_{K|n}(\alpha) = \bar{f}'_{K|n}(\alpha+\epsilon)] \geq (1-g(\epsilon))^K.$$

Because $\lim_{\epsilon\to 0} g(\epsilon) = 0$ and due to the coupling, we have that

$$\mathsf{Pr}[\bar{s}_{K|n}(\alpha) \leq \bar{f}_{K|n}(\alpha)] = \mathsf{Pr}[\bar{s}'_{K|n}(\alpha) \leq \bar{f}'_{K|n}(\alpha)] = \lim_{\epsilon\to 0} \mathsf{Pr}[\bar{s}'_{K|n}(\alpha+\epsilon) \leq \bar{f}'_{K|n}(\alpha+\epsilon)] = \lim_{\epsilon\to 0} \mathsf{Pr}[\bar{s}_{K|n}(\alpha+\epsilon) \leq \bar{f}_{K|n}(\alpha+\epsilon)].$$

This implies that $\mathsf{Pr}[\bar{s}_{K|n}(\alpha) \leq \bar{f}_{K|n}(\alpha)]$ is continuous with respect to α in the interval $(0,1)$.

Because $p_n(\alpha)$ and $\gamma(\alpha)$ are continuous in α in $[0,1]$, and the continuity property is preserved over multiplication and addition, we have that

$$\pi(\alpha) = \sum_{n=1}^{K-1} p_n(\alpha)\mathsf{Pr}[\bar{s}_{K|n}(\alpha) \leq \bar{f}_{K|n}(\alpha)] + \frac{1}{2}(1-\gamma(\alpha))$$

is continuous over $(0,1)$. □

LEMMA 4 **(Set of Equilibria Is Non-Empty and Compact)**. *For any K and $\beta \geq 1$, the set of procedurally rational equilibria, i.e., $\{\alpha : \pi_{K,\beta}(\alpha) = \alpha\}$, is non-empty and compact.*

Proof. Recall from equation (14) in Lemma 3, that $\pi_{K,\beta}(0) = \pi_{K,\beta}(1) = \frac{1}{2}$. This implies that $\pi_{K,\beta}(0) > 0$, while $\pi_{K,\beta}(1) < 1$. By Lemma 3, $\pi_{K,\beta}(\alpha)$ and hence $g_{K,\beta}(\alpha) := \pi_{K,\beta}(\alpha) - \alpha$ is continuous over $[0,1]$. By the intermediate value theorem, there exists at least one $0 < \alpha^* < 1$ such that $\pi_{K,\beta}(\alpha^*) = \alpha^*$.

The set of procedurally rational equilibria is the preimage $g^{-1}_{K,\beta}(\{0\})$. Because $g_{K,\beta}(\alpha)$ is continuous over $[0,1]$, the set of equilibria is therefore closed because the preimage of a closed set (here the singleton $\{0\}$) under a continuous function is closed. The set of equilibria is also bounded as it is contained in $[0,1]$, and it is therefore compact. □

LEMMA 5 **(Limit of Probabilities in K)**. *For each fixed $\alpha \in (\frac{1}{2}, 1)$, and β, we have*

$$\lim_{K\to\infty} \mathsf{Pr}[\bar{s}_{K,\beta}(\alpha) \leq \bar{f}_{K,\beta}(\alpha)|\mathcal{B}] \cdot \gamma_{K,\beta}(\alpha) = 1.$$

Proof. For simplicity of notation, we drop α and β as they are fixed. Let $\bar{s}_K$ ($\bar{f}_K$) be the average waiting time of SF (FS) anecdotes in a sample of K total anecdotes, conditioned on the event $\mathcal{B}$. Also, similar to the proof of Lemma 3, let $\bar{s}_{K|n}$ ($\bar{f}_{K|n}$) be the conditional average waiting time for the SF (FS) anecdotes, *conditioned on drawing exactly n SF anecdotes*. Let $p^S_K(n)$ be the probability of drawing exactly n anecdotes of type SF from a sample of size K. Observe that

$$\begin{aligned}
\mathsf{Pr}[\bar{s}_K \leq \bar{f}_K|\mathcal{B}] \cdot \gamma_{K,\beta}(\alpha) &= \sum_{n=1}^{K-1} p^S_K(n)\mathsf{Pr}[\bar{s}_{K|n} \leq \bar{f}_{K|n}] \\
&= \sum_{n=1}^{K-1} p^S_K(n)(1 - \mathsf{Pr}[\bar{s}_{K|n} > \bar{f}_{K|n}]) \\
&= 1 - p^S_K(0) - p^S_K(K) - \sum_{n=1}^{K-1} p^S_K(n)\mathsf{Pr}[\bar{s}_{K|n} > \bar{f}_{K|n}].
\end{aligned}$$

We have $\lim_{K\to\infty} p_K^S(0) = \lim_{K\to\infty}(1-\alpha)\big[(1-\alpha)/(\beta\alpha+1-\alpha)\big]^{K-1} = 0$, and similarly $\lim_{K\to\infty} p_K^S(K) = \lim_{K\to\infty}\alpha\big[\alpha/(\alpha+\beta(1-\alpha))\big]^{K-1} = 0$. Therefore, we can prove the lemma by showing that

$$\lim_{K\to\infty}\sum_{n=1}^{K-1} p_K^S(n)\mathsf{Pr}[\bar{s}_{K|n} > \bar{f}_{K|n}] = 0.$$

For each of the K draws, the probability that the anecdote comes from route SF is at least $\underline{\rho}_S$ by Assumption 1. The random number of SF anecdotes therefore first-order stochastically dominates a binomial random variable with K trials and success probability $\underline{\rho}_S$. In other words, denoting by $X_K(\underline{\rho}_S)$ a K-trial binomial random variable with success probability $\underline{\rho}_S$, for every $n \geq 0$, we have

$$1-\sum_{k=0}^{n} p_K^S(k) = \mathsf{Pr}[\text{At least } n+1 \ SF \text{ anecdotes}] \geq \mathsf{Pr}[X_K(\underline{\rho}_S) \geq n+1] = 1-\sum_{k=0}^{n}\binom{K}{k}\underline{\rho}_S^k(1-\underline{\rho}_S)^{K-k},$$

which implies that

$$\sum_{k=0}^{n} p_K^S(k) \leq \sum_{k=0}^{n}\binom{K}{k}\underline{\rho}_S^k(1-\underline{\rho}_S)^{K-k} \text{ for each } n \geq 0. \tag{17}$$

By the same assumption, we also have that the probability of the anecdote coming from route FS on each draw is at least $\underline{\rho}_F$. We thus have

$$1-\sum_{k=K-n}^{K} p_K^S(k) = \mathsf{Pr}[\text{At least } n+1 \ FS \text{ anecdotes}] \geq 1-\sum_{k=0}^{n}\binom{K}{k}\underline{\rho}_F^k(1-\underline{\rho}_F)^{K-k},$$

which implies that

$$\sum_{k=K-n}^{K} p_K^S(k) \leq \sum_{k=0}^{n}\binom{K}{k}\underline{\rho}_F^k(1-\underline{\rho}_F)^{K-k} \text{ for each } n \geq 0. \tag{18}$$

For any fixed n, the right-hand side of both inequalities (17) and (18) goes to 0 as K goes to infinity. So, we have that for each fixed n and any $\epsilon > 0$, there exists $K(n)$ such that for all $K \geq K(n)$,

$$\sum_{k=0}^{n} p_K^S(k) + \sum_{k=K-n}^{K} p_K^S(k) < \frac{\epsilon}{3}. \tag{19}$$

Let s^* (f^*) be the true expected system time for route SF (FS), given the fixed SF fraction $\alpha \in (1/2, 1)$. By Lemma 1, we have

$$s^* = \frac{1}{\alpha}\int_0^{\alpha} \mathcal{S}(\alpha; y_S)dy_S \leq \frac{1}{\alpha}\int_0^{\alpha}\frac{y_S+1-\alpha}{\mu_S}dy_S = \frac{1-\alpha/2}{\mu_S}; \tag{20}$$

and by Lemma 2, the expected FS system time f^* is equal to

$$f^* = \frac{1}{1-\alpha}\int_0^{1-\alpha} \mathcal{F}(\alpha; y_F)dy_F = \frac{1}{1-\alpha}\int_0^{1-\alpha}\frac{\alpha+y_F}{\mu_S}dy_F = \frac{\alpha/2+1/2}{\mu_S}. \tag{21}$$

Note that $\alpha \in (1/2, 1)$ implies that $1-\alpha/2 < \alpha/2+1/2$. Combining this with equations (20) and (21), we have that $d := f^* - s^* > 0$.

By the weak law of large numbers (see, e.g., Grimmett and Stirzaker 2020, Section 7.4), for any $\epsilon > 0$, there exists N_0 such that for all K and n with $n \geq N_0$ and $K-n \geq N_0$, we have that

$$\mathsf{Pr}[\bar{s}_{K|n} \geq s^* + \frac{d}{2}] < \frac{\epsilon}{3}, \text{ and } \mathsf{Pr}[\bar{f}_{K|n} \leq f^* - \frac{d}{2}] < \frac{\epsilon}{3},$$

which in turn implies that

$$\mathsf{Pr}[\bar{s}_{K|n} > \bar{f}_{K|n}] < \frac{2\epsilon}{3}. \tag{22}$$

Now, for any $\epsilon > 0$, choose N_0 so that inequality (22) is satisfied for all $n \geq N_0$, and pick $K_0 = K(N_0)$ such that inequality (19) holds for all $K \geq K_0$. Then, we have

$$\begin{aligned}
\sum_{n=1}^{K-1} p_K^S(n)\mathsf{Pr}[\bar{s}_{K|n} > \bar{f}_{K|n}] &\leq \sum_{n=0}^{N_0} p_K^S(n) + \sum_{n=K-N_0}^{K} p_K^S(n) + \sum_{n=N_0}^{K-N_0} p_K^S(n)\mathsf{Pr}[\bar{s}_{K|n} > \bar{f}_{K|n}] \\
\text{By (19),} \qquad &< \frac{\epsilon}{3} + \sum_{n=N_0}^{K-N_0} p_K^S(n)\mathsf{Pr}[\bar{s}_{K|n} > \bar{f}_{K|n}] \\
\text{By (22),} \qquad &\leq \frac{\epsilon}{3} + \frac{2\epsilon}{3}\sum_{n=N_0}^{K-N_0} p_K^S(n) \\
&\leq \epsilon.
\end{aligned}$$

This establishes

$$\lim_{K\to\infty} \sum_{n=1}^{K-1} p_K^S(n)\mathsf{Pr}[\bar{s}_{K|n} > \bar{f}_{K|n}] = 0,$$

which completes the proof of the lemma. □

Proof of Proposition 1. By Lemma 5, for any $\alpha \in (1/2, 1)$ and $\beta \geq 1$, there exists K_0 such that for $K \geq K_0$,

$$\mathsf{Pr}[\bar{s}_{K,\beta}(\alpha) \leq \bar{f}_{K,\beta}(\alpha)|\mathcal{B}] \cdot \gamma_{K,\beta}(\alpha) > \alpha,$$

which implies that $\pi_{K,\beta}(\alpha) > \alpha$. As $\pi_{K,\beta}(1) = 1/2$ and $\pi_{K,\beta}$ is continuous, there exists an equilibrium that is strictly larger than α. Because this is true for any $1/2 \leq \alpha < 1$, it implies that $\liminf_{K\to\infty} \alpha_\beta \geq 1$. Also, by definition, the largest equilibrium is no larger than 1, thus $\limsup_{K\to\infty} \alpha_\beta \leq 1$, implying that $\lim_{K\to\infty} \alpha_\beta = 1$. This establishes that $\lim_{K\to\infty} \alpha_\beta = 1$. □

Proof of Corollary 1. By equations (20) and (21), we have that for $\alpha \in (1/2, 1)$ the expected system time of SF is strictly less than the expected system time of FS. Thus, $\pi^*(\alpha) = 1$ for $\alpha \in (1/2, 1)$.

Next, for $\alpha \in (1/2, 1)$, $\pi_{K,\beta}(\alpha) \leq 1$ for any K, β, and by Lemma 5,

$$\lim_{K\to\infty} \pi_{K,\beta}(\alpha) \geq \lim_{K\to\infty} \mathsf{Pr}[\bar{s}_{K,\beta}(\alpha) \leq \bar{f}_{K,\beta}(\alpha)|\mathcal{B}] \cdot \gamma_{K,\beta}(\alpha) = 1.$$

Therefore, we have that for $\alpha \in (1/2, 1)$

$$\lim_{K\to\infty} \pi_{K,\beta}(\alpha) = \pi^*(\alpha) = 1.$$

□

Proof of Theorem 1. Throughout the proof, we consider a fixed sample size $K \geq 2$. Consider $\alpha \in (\mu_S/\mu_F, 1)$. Conditional on a customer's sample containing both types of anecdotes: (i) Lemma 1 implies $\sup\{\text{supp}(\bar{s}_{K,\beta}(\alpha))\} = \mathcal{S}(\alpha; \alpha) = \alpha/\mu_S$; and (ii) Lemma 2 implies $\inf\{\text{supp}(\bar{f}_{K,\beta}(\alpha))\} = \alpha/\mu_S$. We therefore have

$$\mathsf{Pr}[\bar{s}_{K,\beta}(\alpha) \leq \bar{f}_{K,\beta}(\alpha)|\mathcal{B}] = 1.$$

By equation (3), we then have

$$\pi_{K,\beta}(\alpha) = \gamma_{K,\beta}(\alpha) + \frac{1}{2}(1 - \gamma_{K,\beta}(\alpha)) = \frac{1}{2}\gamma_{K,\beta}(\alpha) + \frac{1}{2}. \tag{23}$$

We also have

$$\gamma_{K,\beta}(\alpha) = \sum_{n=1}^{K-1} p_K^S(n) = 1 - p_K^S(0) - p_K^S(K), \tag{24}$$

where

$$p_K^S(0) = \alpha\Big(1 - \frac{\beta\alpha}{\beta\alpha + (1-\alpha)}\Big)^{K-1} \quad \text{and} \quad p_K^S(K) = (1-\alpha)\Big(1 - \frac{\beta(1-\alpha)}{\alpha + \beta(1-\alpha)}\Big)^{K-1},$$

by equations (2) and (1), respectively.

We have

$$\frac{\partial p_K^S(0)}{\partial \beta} = \alpha(K-1)\Big(1 - \frac{\beta\alpha}{\beta\alpha + (1-\alpha)}\Big)^{K-2}\Big(-\frac{\partial(\beta\alpha/(\beta\alpha + (1-\alpha))}{\partial\beta}\Big).$$

The first three terms of this expression are positive, so the sign is determined by the sign of the last term. For this term, we can write

$$\frac{\partial(\beta\alpha/(\beta\alpha + (1-\alpha))}{\partial\beta} = \frac{\alpha}{\beta\alpha + (1-\alpha)}\Big(1 - \frac{\beta\alpha}{\beta\alpha + (1-\alpha)}\Big) > 0,$$

implying that $\partial p_K^S(0)/\partial\beta < 0$. Analogous manipulations imply that also $\partial p_K^S(K)/\partial\beta < 0$.

We can then deduce from equation (24) that the probability $\gamma_{K,\beta}(\alpha)$ is increasing in β for given α, which by equation (23) implies that $\pi_{K,\beta}(\alpha)$ is increasing in β. We also have $\lim_{\beta\to\infty} p_K^S(0) = \lim_{\beta\to\infty} p_K^S(K) = 0$, which implies that $\lim_{\beta\to\infty} \gamma_{K,\beta}(\alpha) = 1$. In turn, we have

$$\lim_{\beta\to\infty} \pi_{K,\beta}(\alpha) = 1. \tag{25}$$

For any $\alpha \in (\mu_S/\mu_F, 1)$, equation (25) implies that for all β above some threshold M_0, we must have $\pi_{K,\beta}(\alpha) > \alpha$. As $\pi_{K,\beta}(1) = 1/2$ and $\pi_{K,\beta}$ is continuous, for all $\beta > M_0$ there exists an equilibrium that is strictly larger than α. Because this is true for any $1/2 \le \alpha < 1$, it implies that $\liminf_{\beta\to\infty} \alpha_{K,\beta} \ge 1$. Also, by definition, the largest equilibrium is no larger than 1, thus $\limsup_{\beta\to\infty} \alpha_{K,\beta} \le 1$, implying that $\lim_{\beta\to\infty} \alpha_{K,\beta} = 1$. This establishes part (i).

To prove part (ii), note that $\lim_{\beta\to\infty} \alpha_{K,\beta} = 1$ by definition implies that for fixed $\alpha \in (\mu_S/\mu_F, 1)$, there exists $\hat{\beta}(\alpha)$ such that $\alpha_{K,\beta} > \alpha$ for all $\beta > \hat{\beta}(\alpha)$. Thus, there exists $\hat{\beta}$—e.g., $\hat{\beta}(\mu_S/\mu_F + \epsilon)$ for small $\epsilon > 0$—such that for all $\beta > \hat{\beta}$, the largest equilibrium satisfies $\mu_S/\mu_F < \alpha_{K,\beta} < 1$. We now consider the largest equilibrium $\alpha_{K,\beta}$ for some $\beta > \hat{\beta}$. We have $\mu_S/\mu_F < \alpha_{K,\beta} < 1$, which as shown above means that $\pi_{K,\beta}(\alpha_{K,\beta})$ is increasing in β. Therefore, for $\epsilon > 0$ we have $\pi_{K,\beta+\epsilon}(\alpha_{K,\beta}) > \alpha_{K,\beta}$, so by the intermediate value theorem and the fact that $\pi_{K,\beta}(1) = 1/2$ (see equation 14), we have $\alpha_{K,\beta+\epsilon} > \alpha_{K,\beta}$. We conclude that the largest equilibrium $\alpha_{K,\beta}$ is increasing in β for $\beta > \hat{\beta}$. □

E. Proofs for Section 4.2

To understand the procedurally rational response function π_β (note that we drop the dependence on K in this section as it is here fixed at 2) requires first deriving $\phi(\alpha)$, which is the probability—conditional on receiving exactly one anecdote from each route— that the SF anecdote will reflect a shorter system time than the FS anecdote. This we accomplish in Lemma 6. We then derive structural properties of ϕ in Lemmas 7 and 8. Then, in Lemma 9, we prove that π_β converges to ϕ in the limit of β. Lemma 10 shows that the procedurally rational response function π_β is increasing in β for given α with $\phi(\alpha) > 1/2$, while Lemma 11 shows that the largest equilibrium is always larger than 1/2. Finally, Lemma 12 establishes that the largest equilibrium increases with β and converges to herding as $\beta \to \infty$. Note that Lemma 12 generalizes Theorem 1 for the special case of $K = 2$ in that (i) it includes the case with $\mu_S = \mu_F$ and (ii) the increasing property holds for all $\beta > 1$.

Using these auxiliary results, we then prove Proposition 2, which derives the unique equilibrium for the case of $\beta = 1$. After that, we give the proof of Proposition 3, that there is exactly one equilibrium between 1/2 and 1 under almost any parameters.

LEMMA 6 **(Characterization of ϕ)**. *For a customer that receives one sample from each route, we have*

$$\phi(\alpha) = \mathsf{Pr}[\hat{s} \le \hat{f}] = \begin{cases} 1 - \dfrac{\mu_S}{\mu_F} + \dfrac{\alpha}{1-\alpha}\Big(1 - \dfrac{\mu_S}{2\mu_F}\Big) & \text{if } 0 < \alpha < \dfrac{\mu_S}{\mu_S + \mu_F}, \\ 1 + \dfrac{1}{1-\alpha} - \dfrac{\mu_S^2 + \alpha^2\mu_F^2}{2\alpha(1-\alpha)\mu_S\mu_F} & \text{if } \dfrac{\mu_S}{\mu_S + \mu_F} \le \alpha < \dfrac{\mu_S}{\mu_F}, \\ 1 & \text{if } \dfrac{\mu_S}{\mu_F} \le \alpha < 1. \end{cases} \tag{26}$$

In addition, $\phi(\cdot)$ is continuous over $(0,1)$.

Proof. Let $\hat{y}_S$ and $\hat{y}_F$ be the position in the queue of the randomly selected SF and FS anecdotes in a sample, where $\hat{y}_S$ and $\hat{y}_F$ are uniform draws from the intervals $[0,\alpha]$ and $[0,1-\alpha]$, respectively. The revealed system times, denoted by $\hat{s}$ and $\hat{f}$, are related to the positions $\hat{y}_S$ and $\hat{y}_F$ of the randomly drawn anecdotes by the expressions in Lemmas 1 and 2. Denoted by $\phi(\alpha)$, the response probability that a customer who receives one anecdote from each route chooses route SF is thus the probability that $\hat{s} \le \hat{f}$. We proceed by cases.

Case 1: $0 < \alpha < \mu_S/(\mu_S + \mu_F)$. Straightforward algebra reveals that $\alpha < \mu_S/(\mu_S + \mu_F) < \mu_S/\mu_F$ implies $\alpha < \mu_S(1-\alpha)/(\mu_F - \mu_S)$. Thus, because $y_S \le \alpha$, we can ignore the second piece of the function $\mathcal{S}(\cdot)$ given in equation (11). A customer receiving anecdotes from both routes will choose route SF if and only if her random draws satisfy $\hat{s} \le \hat{f}$, that is, if and only if

$$\frac{\hat{y}_S + 1 - \alpha}{\mu_F} \le \frac{\alpha + \hat{y}_F}{\mu_S} \iff \frac{\mu_S}{\mu_F}\big(\hat{y}_S + 1 - \alpha\big) - \alpha \le \hat{y}_F.$$

Therefore, the probability that a customer with anecdotes of both types chooses route SF is given by

$$\mathsf{Pr}[\hat{s} \le \hat{f}] = \mathsf{Pr}\Big[\frac{\mu_S}{\mu_F}\Big(\hat{y}_S + 1 - \alpha - \frac{\alpha\mu_F}{\mu_S}\Big) \le \hat{y}_F\Big]. \tag{27}$$

Because the position of the randomly drawn FS customer is uniformly distributed within the station F buffer, the random variable $\hat{y}_F$ has a uniform distribution on the interval $[0,1-\alpha]$. Furthermore, $\alpha < \mu_S/(\mu_S + \mu_F)$ implies

$$\frac{\mu_S}{\mu_F}\big(1 - \alpha\big) - \alpha > 0.$$

Therefore, conditional on a value of $\hat{y}_S$, we have

$$\mathsf{Pr}\big[\hat{s} \le \hat{f} \big| \hat{y}_S\big] = 1 - \frac{\mu_S\big(\hat{y}_S + 1 - \alpha\big) - \alpha\mu_F}{\mu_F(1-\alpha)}. \tag{28}$$

Similar to $\hat{y}_F$, the random variable $\hat{y}_S$ has a uniform distribution on $[0,\alpha]$. It therefore has density function $g(x)$, where

$$g(x) = \begin{cases} \dfrac{1}{\alpha} & \text{if } 0 \le x \le \alpha, \\ 0 & \text{otherwise.} \end{cases}$$

Integrating over this density function, we compute $\phi(\alpha)$, the probability that a customer chooses route SF conditional on receiving anecdotes of both types, by

$$\begin{aligned} \phi(\alpha) &= \mathsf{Pr}[\hat{s} \le \hat{f}] \\ &= \int_0^\alpha \mathsf{Pr}\big[\hat{s} \le \hat{f} \big| \hat{y}_S = x\big] f(x) \mathsf{d}x \\ &= \frac{1}{\alpha}\int_0^\alpha \Big[1 - \frac{\mu_S(x + 1 - \alpha) - \alpha\mu_F}{\mu_F(1-\alpha)}\Big] \mathsf{d}x \\ &= 1 - \frac{\mu_S}{\mu_F} + \frac{\alpha}{1-\alpha}\Big(1 - \frac{\mu_S}{2\mu_F}\Big). \end{aligned}$$

Case 2: $\mu_S/(\mu_S+\mu_F) \le \alpha \le \mu_S/\mu_F$. The probability that a customer with anecdotes of both types will choose route SF is again the probability that $\hat{s} \le \hat{f}$, and it takes the same form as equation (27). If the left-hand side (LHS) of the inequality inside the brackets in equation (27) is negative, then the probability is necessarily equal to 1 because $\hat{y}_F$ takes only nonnegative values. If the LHS of the inequality is nonnegative, then the probability is equal to the expression given in equation (28). So, conditional on a value of $\hat{y}_S$ such that $0 \le \hat{y}_S \le \alpha$, we have

$$\mathsf{Pr}[\hat{s} \le \hat{f} | \hat{y}_S] = \begin{cases} 1 & \text{if } \hat{y}_S + 1 - \alpha - \dfrac{\alpha\mu_F}{\mu_S} \le 0, \\ 1 - \dfrac{\mu_S(\hat{y}_S + 1 - \alpha) - \alpha\mu_F}{\mu_F(1-\alpha)} & \text{otherwise.} \end{cases}$$

The remainder of the proof is exactly analogous to the proof of Case 1, with the end result being the expression in the second piece of equation (26).

Case 3: $\mu_S/\mu_F < \alpha < 1$. If $\mu_S/\mu_F < \alpha < 1$ (which we note is possible only if $\mu_S < \mu_F$), then for $0 \le y_S \le \alpha$, we have

$$\mathcal{S}(y_S) \le \mathcal{S}(\alpha) = \frac{\alpha}{\mu_S},$$

and

$$\mathcal{F}(y_F) = \frac{\alpha + y_F}{\mu_S} \ge \frac{\alpha}{\mu_S} = \mathcal{S}(\alpha) \quad \text{for } 0 \le y_F \le 1 - \alpha.$$

Thus, for any possible draw $\hat{y}_S$ for route SF and any possible draw $\hat{y}_F$ for route FS, we have $\mathcal{S}(\hat{y}_S) \le \mathcal{F}(\hat{y}_F)$. Thus, for $\mu_S/\mu_F < \alpha < 1$, we have

$$\phi(\alpha) = \mathsf{Pr}[\hat{s} \le \hat{f}] = 1.$$

Combining Cases 1-3, we conclude equation (26). Continuity of each of the three pieces is immediate, and continuity at the boundaries is readily verified by substitution of the endpoints into the expressions for the intersecting pieces of the function. □

LEMMA 7 **(Concave Interval of ϕ)**. *The function $\phi(\alpha)$ is increasing and strictly concave in α on the interval $[\mu_S/(\mu_S+\mu_F), \mu_S/\mu_F]$.*

Proof. Differentiating ϕ at the interval of interest twice, we get

$$\phi''(\alpha) = \frac{(\mu_F - \mu_S)^2}{\mu_S\mu_F(\alpha-1)^3} - \frac{\mu_S}{\mu_F\alpha^3}. \tag{29}$$

The first term in equation (29) is nonpositive because $0 < \alpha < 1$ in the range we are considering. The second term in equation (29) is strictly positive because μ_S, μ_F, and α are all strictly positive. Thus, equation (29) entails subtracting a positive number from a nonpositive one, and we conclude

$$\phi''(\alpha) = \frac{(\mu_F - \mu_S)^2}{\mu_S\mu_F(\alpha-1)^3} - \frac{\mu_S}{\mu_F\alpha^3} < 0.$$

Therefore, the function $\phi(\alpha)$ is strictly concave for $\mu_S/(\mu_S+\mu_F) \le \alpha \le \mu_S/\mu_F$ by the second-order condition for concavity. Now, the fact that the second derivative $\phi''(\alpha) < 0$ implies that the first derivative is strictly decreasing. Evaluating the first derivative gives

$$\phi'(\alpha) = \frac{-(\mu_F - \mu_S)^2}{2\mu_S\mu_F(1-\alpha)^2} + \frac{\mu_S}{2\mu_F\alpha^2}. \tag{30}$$

Substituting $\alpha = \mu_S/\mu_F$ gives

$$\phi'\left(\frac{\mu_S}{\mu_F}\right) = 0,$$

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

$$\phi'(\alpha) \geq 0 \quad \text{for } \frac{\mu_S}{\mu_S+\mu_F} \leq \alpha \leq \frac{\mu_S}{\mu_F}.$$

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

LEMMA 8 **(Property of ϕ)**. *Suppose that $\mu_S < \mu_F$. For $\mu_S/(\mu_S+\mu_F) \leq \alpha < 1$, we have $\phi(\alpha) > \alpha$.*

Proof. Substituting into equation (26), we have

$$\phi\Big(\frac{\mu_S}{\mu_S+\mu_F}\Big) = 1 - \frac{\mu_S^2}{2\mu_F^2}.$$

Then, $\mu_S < \mu_F$ implies

$$\begin{aligned} & \mu_S^3 + \mu_S^2\mu_F < 2\mu_F^3 \\ \iff & 2\mu_F^2\mu_S + \mu_S^2(\mu_S+\mu_F) < 2\mu_F^2(\mu_S+\mu_F) \\ \iff & \frac{\mu_S}{\mu_S+\mu_F} < 1 - \frac{\mu_S^2}{2\mu_F^2} = \phi\Big(\frac{\mu_S}{\mu_S+\mu_F}\Big). \end{aligned} \tag{31}$$

Again substituting into equation (26), we get

$$\phi\Big(\frac{\mu_S}{\mu_F}\Big) = 1 > \frac{\mu_S}{\mu_F}. \tag{32}$$

By Lemma 7, the function $\phi(\alpha)$ is concave for $\mu_S/(\mu_S+\mu_F) \leq \alpha \leq \mu_S/\mu_F$. By the definition of concavity—see, e.g., Boyd and Vandenberghe (2004, Section 3.1)—for any $\theta \in [0,1]$, we then have

$$\phi(\theta x + (1-\theta)x') \geq \theta\phi(x) + (1-\theta)\phi(x') \quad \text{for } x, x' \in \Big[\frac{\mu_S}{\mu_S+\mu_F}, \frac{\mu_S}{\mu_F}\Big]. \tag{33}$$

We can express any $\alpha \in [\mu_S/(\mu_S+\mu_F), \mu_S/\mu_F]$ as

$$\alpha = \theta\Big(\frac{\mu_S}{\mu_S+\mu_F}\Big) + (1-\theta)\Big(\frac{\mu_S}{\mu_F}\Big),$$

where α ranges from $\mu_S/(\mu_S+\mu_F)$ to μ_S/μ_F as θ ranges from 0 to 1. Taking $x = \mu_S/(\mu_S+\mu_F)$ and $x' = \mu_S/\mu_F$ in equation (33), we get

$$\begin{aligned} \phi(\alpha) = \phi\Big(\theta\Big(\frac{\mu_S}{\mu_S+\mu_F}\Big) + (1-\theta)\Big(\frac{\mu_S}{\mu_F}\Big)\Big) &\geq \theta\phi\Big(\frac{\mu_S}{\mu_S+\mu_F}\Big) + (1-\theta)\phi\Big(\frac{\mu_S}{\mu_F}\Big) \\ &> \theta\Big(\frac{\mu_S}{\mu_S+\mu_F}\Big) + (1-\theta)\Big(\frac{\mu_S}{\mu_F}\Big) = \alpha, \end{aligned}$$

where the strict inequality holds because $\phi(\mu_S/(\mu_S+\mu_F)) > \mu_S/(\mu_S+\mu_F)$ and $\phi(\mu_S/\mu_F) > \mu_S/\mu_F$, by equations (31) and (32), respectively. We conclude that $\phi(\alpha) > \alpha$ for all $\mu_S/(\mu_S+\mu_F) \leq \alpha \leq \mu_S/\mu_F$.

Finally, for $\mu_S/\mu_F < \alpha < 1$, we have $\phi(\alpha) = 1 > \alpha$. □

LEMMA 9 **(Limit of π_β is ϕ)**. *For $0 < \alpha < 1$, we have*

$$\lim_{\beta\to\infty} \pi_\beta(\alpha) = \phi(\alpha).$$

Proof. First, taking the limit of γ_β in equation (4), we get

$$\begin{aligned} \lim_{\beta\to\infty} \gamma_\beta(\alpha) &= \lim_{\beta\to\infty} \left(\frac{\beta\alpha(1-\alpha)}{\alpha+\beta(1-\alpha)} + \frac{\beta\alpha(1-\alpha)}{\beta\alpha+(1-\alpha)} \right) \\ &= \frac{\alpha(1-\alpha)}{1-\alpha} + \frac{\alpha(1-\alpha)}{\alpha} \\ &= 1. \end{aligned} \tag{34}$$

By the definition of π_β in equation (5), we then have

$$\begin{aligned}\lim_{\beta\to\infty}\pi_\beta(\alpha) &= \lim_{\beta\to\infty}\Big(\gamma_\beta(\alpha)\phi(\alpha)+\frac{1}{2}\big(1-\gamma_\beta(\alpha)\big)\Big)\\ &= \frac{1}{2}+\Big(\phi(\alpha)-\frac{1}{2}\Big)\lim_{\beta\to\infty}\gamma_\beta(\alpha)\\ &= \phi(\alpha),\end{aligned}$$

where the last equality follows from equation (34), completing the proof. □

LEMMA 10 **(Slope of Response π_β in β)**. *For given $0<\alpha<1$, the response function $\pi_\beta(\alpha)$ is strictly increasing in β if and only if $\phi(\alpha)>1/2$.*

Proof. Differentiating the probability γ_β with respect to β gives

$$\frac{\partial\gamma_\beta}{\partial\beta}=\frac{\alpha(1-\alpha)^2}{(1-\alpha+\alpha\beta)^2}+\frac{(1-\alpha)\alpha^2}{(\alpha+\beta-\alpha\beta)^2}>0.$$

Because $\phi(\alpha)$ is constant in β, if we then differentiate the response function π_β with respect to β, from equation (5), we get

$$\frac{\partial\pi_\beta}{\partial\beta}=\frac{\partial\gamma_\beta}{\partial\beta}\Big(\phi(\alpha)-\frac{1}{2}\Big).$$

Therefore, we have that $\pi_\beta(\alpha)$ is strictly increasing in β for given α if and only if $\phi(\alpha)>1/2$. □

LEMMA 11 **(Largest Equilibrium Is Above 1/2)**. *For any discernibility parameter β, we have $\alpha_\beta\geq 1/2$ and $\phi(\alpha_\beta)\geq 1/2$. The inequalities are strict if at least one of $\mu_S<\mu_F$ or $\beta>1$ holds.*

Proof. Consider $\alpha=1/2$, and note that it can be shown that $\phi(\alpha)>1/2$ if and only if $\alpha>(2\mu_S-\mu_F)/(\mu_S+\mu_F)$. For $\mu_S<\mu_F$, we have

$$\frac{1}{2}=\frac{\mu_S}{2\mu_S}>\frac{2\mu_S-\mu_F}{\mu_S+\mu_F}, \tag{35}$$

which implies that $\phi(\alpha)>1/2$ for all $\alpha\geq 1/2$. Combining this observation with equation (5), for $\beta>0$, we can then write

$$\begin{aligned}\pi_\beta\Big(\frac{1}{2}\Big) &= \gamma_\beta\Big(\frac{1}{2}\Big)\phi\Big(\frac{1}{2}\Big)+\frac{1}{2}-\frac{1}{2}\gamma_\beta\Big(\frac{1}{2}\Big)\\ &= \frac{1}{2}+\gamma_\beta\Big(\frac{1}{2}\Big)\Big(\phi\Big(\frac{1}{2}\Big)-\frac{1}{2}\Big)\\ &> \frac{1}{2},\end{aligned} \tag{36}$$

where the inequality follows because we have $\phi(1/2)>1/2$ and $\gamma_\beta(1/2)>0$, the latter of which can be seen by inspection of equation (4). Hence, we have $\pi_\beta(1/2)>1/2$, and we also have $\pi_\beta(1)<1$, as shown in the proof of Lemma 4. Therefore, by the intermediate value theorem, at least one equilibrium α^* exists such that $1/2<\alpha^*<1$, implying that the largest equilibrium α_β is strictly larger than $1/2$. By earlier arguments, this also implies that $\phi(\alpha_\beta)>1/2$.

For $\mu_S=\mu_F$, substitution into equation (26) reveals that $\phi(1/2)=1/2$. Equation (36) then implies that $\pi_\beta(1/2)=1/2$, and therefore $\alpha_\beta\geq 1/2$. By Lemma 7, we then have $\phi(\alpha_\beta)\geq\phi(1/2)=1/2$. Moreover, if $\mu_S=\mu_F$ and $\beta>1$, then differentiating the second piece of equation (26) gives $\pi'_\beta(1/2)=2\beta/(\beta+1)>1$. This implies that $\pi_\beta(\alpha)-\alpha$ is strictly increasing at $\alpha=1/2$, so because $\pi_\beta(1/2)-1/2=0$, there exists $\epsilon>0$ such that $\pi_\beta(1/2+\epsilon)>1/2+\epsilon$. Therefore, there is an equilibrium α_β strictly larger than $1/2$, and because $\alpha_\beta>1/2=(2\mu_S-\mu_F)/(\mu_S+\mu_F)$, we also have $\phi(\alpha_\beta)>1/2$. □

LEMMA 12 **(Properties of the Largest Equilibrium for $K=2$)**. *Under the procedurally rational model with $K=2$, the largest equilibrium, α_β, is increasing in β for $\beta \in (1,\infty)$. Moreover, α_β converges to herding as $\beta \to \infty$, i.e., we have* $\lim_{\beta\to\infty} \alpha_\beta = 1$.

Proof. By Lemma 11, for any fixed $\beta > 1$, we have that $\alpha_\beta > 1/2$ and $\phi(\alpha_\beta) > 1/2$. Then Lemma 10 implies that $\pi_\beta(\alpha_\beta)$ is strictly increasing in β, implying that for any $\epsilon > 0$,

$$\pi_{\beta+\epsilon}(\alpha_\beta) > \alpha_\beta.$$

Combining this with $\pi_{\beta+\epsilon}(1) = 1/2$, by the intermediate value theorem, there exists an equilibrium strictly larger than α_β for discernibility parameter $\beta+\epsilon$. Therefore, the largest equilibrium $\alpha_{\beta+\epsilon}$ satisfies $\alpha_\beta < \alpha_{\beta+\epsilon}$, and we conclude that the largest equilibrium is increasing in β.

Next, we prove $\lim_{\beta\to\infty} \alpha_\beta = 1$. For $\mu_S < \mu_F$, by Lemma 8, because $\mu_S/(\mu_S+\mu_F) < 1/2$, we have

$$\phi(\alpha) > \alpha \text{ for } \alpha \geq \frac{1}{2}. \tag{37}$$

For $\mu_S = \mu_F$, by Proposition 6, we have

$$\phi(\alpha) = \frac{3}{2} - \frac{1}{2\alpha} > \alpha \text{ for } \alpha > \frac{1}{2}. \tag{38}$$

Thus, for any $\mu_S \leq \mu_F$, we have that $\phi(\alpha) > \alpha$ for all $\alpha > 1/2$.

By Lemma 9, this implies that there exists some M_0 such that for any $\beta > M_0$ and $1/2 \leq \alpha < 1$

$$\pi_\beta(\alpha) > \alpha. \tag{39}$$

As $\pi_\beta(1) = 1/2$ and π_β is continuous, for all $\beta > M_0$ there exists an equilibrium that is strictly larger than α. Because this is true for any $1/2 \leq \alpha < 1$, it implies that $\liminf_{\beta\to\infty} \alpha_\beta \geq 1$. Also, by definition, the largest equilibrium is no larger than 1, thus $\limsup_{\beta\to\infty} \alpha_\beta \leq 1$, implying that $\lim_{\beta\to\infty} \alpha_\beta = 1$. □

With the above lemmas, we now prove Propositions 2 and 12, as well as Proposition 3.

Proof of Proposition 2. **Case 1: $\mu_S/\mu_F \leq 1/\sqrt{2}$.** In the rightmost interval of ϕ, i.e., for $\mu_S/\mu_F < \alpha \leq 1$, we have $\phi(\alpha) = 1$. Therefore, we can write

$$\pi_1(\alpha) = \phi(\alpha)\gamma_1(\alpha) + \frac{1-\gamma_1(\alpha)}{2} = \frac{1}{2} + \alpha - \alpha^2.$$

Thus, in the interval $[\mu_S/\mu_F, 1]$, the only candidate solution to the fixed-point equation $\pi_1(\alpha) = \alpha$ is

$$\alpha^* = \frac{1}{\sqrt{2}}.$$

The assumption of Case 1 implies that $\mu_S/\mu_F \leq \alpha^*$, so α^* falls into the corresponding interval of the function. We conclude that α^* is an equilibrium, and it is therefore the largest equilibrium because it is the only equilibrium in the rightmost interval of α.

Case 2: $\mu_S/\mu_F > 1/\sqrt{2}$. In this case, by the reasoning above, there is no equilibrium in the rightmost interval $[\mu_S/\mu_F, 1]$. By Lemma 11, the largest equilibrium falls into the interval $[1/2, 1)$ (the interval is open on the left if the service rates are nonidentical), so in this case the largest equilibrium must fall in $[1/2, \mu_S/\mu_F)$. In this interval, the functional form that governs π_1 is that which covers the interval $[\mu_S/(\mu_S+\mu_F), \mu_S/\mu_F]$. Applying the expressions for γ_β and π_β in equations (4) and (5), along with the expression for ϕ from Proposition 6, in this interval, we can express the function $\pi_1(\alpha)$ by

$$\pi_1(\alpha) = \frac{1}{2} - \frac{\mu_S}{\mu_F} + 3\alpha - \alpha^2\Big(1 + \frac{\mu_F}{\mu_S}\Big).$$

This is a quadratic equation with two real roots

$$\frac{\mu_S}{\mu_S+\mu_F}\left(1\pm\sqrt{\frac{(\mu_F-\mu_S)(2\mu_S+\mu_F)}{2\mu_S\mu_F}}\right).$$

The smaller root cannot be an equilibrium because it is less than $\mu_S/(\mu_S+\mu_F)$ and therefore not in the interval in which the corresponding expression governs π_1. We denote the larger root by α'. Inspection reveals that $\alpha' \geq \mu_S/(\mu_S+\mu_F)$, and it can be shown that

$$\alpha' < \frac{\mu_S}{\mu_F} \iff \frac{\mu_S}{\mu_F} > \frac{1}{\sqrt{2}}.$$

In other words, the condition for α' to fall into the appropriate interval (and therefore constitute an equilibrium) is exactly the converse of the condition for α^* to be an equilibrium. We conclude equation (6) and that there is exactly one equilibrium in the interval $[\mu_S/(\mu_S+\mu_F), 1]$.

Finally, in the interval $[0, \mu_S/(\mu_S+\mu_F)]$, the candidates for equilibria can be found by solving another quadratic fixed-point equation, arrived at by analogous steps to the above for the first interval of the function ϕ. The solutions to this equation are given by

$$\frac{\mu_S}{\mu_S+\mu_F} \pm \frac{\sqrt{4\mu_S^2-2\mu_S\mu_F-2\mu_F^2}}{2(\mu_S+\mu_F)}.$$

The quantity under the radical is strictly negative for $\mu_S < \mu_F$, in which case there are no real solutions to the fixed-point equation and no equilibrium in the interval $[0, \mu_S/(\mu_S+\mu_F)]$. If $\mu_S = \mu_F$, then the radical is zero and both roots are equal to $1/2$, which is also equal to α' in this case. In both cases, we do not have an *additional* equilibrium in this interval. We conclude that there is exactly one equilibrium on $[0,1]$, given by equation (6). □

Proof of Proposition 3. First, we clarify that $\pi'_\beta(\alpha)$ denotes the derivative of π with respect to α for given β, and similarly $\gamma'_\beta(\alpha)$ the derivative of γ_β with respect to α for given β. We can rearrange equation (5) to give

$$\pi_\beta(\alpha) = \gamma_\beta(\alpha)\left(\phi(\alpha)-\frac{1}{2}\right)+\frac{1}{2}.$$

Differentiating with respect to α yields

$$\pi'_\beta(\alpha) = \gamma'_\beta(\alpha)\left(\phi(\alpha)-\frac{1}{2}\right)+\gamma_\beta(\alpha)\phi'(\alpha). \tag{40}$$

We will treat separately the two terms on the RHS of equation (40). The derivative of γ_β is given by

$$\gamma'_\beta(\alpha) = \frac{(1-2\alpha)\beta^2(1+\beta)}{\left((\beta-1)^2(\alpha-\alpha^2)+\beta\right)^2}.$$

The denominator is strictly positive, and the numerator is nonpositive for $\alpha \geq 1/2$, implying that $\gamma'_\beta(\alpha)$ is nonpositive for $\alpha \geq 1/2$. Moreover, the numerator is decreasing in α for $\alpha \geq 1/2$, and the denominator is decreasing in α over the same interval by our assumption that $\beta \geq 1$, implying that $\gamma'_\beta(\alpha)$ is strictly decreasing in α because it is negative and increasing in magnitude with α. As noted in the proof of Lemma 11, we have $\phi(\alpha) > 1/2$ for $\alpha > 1/2$, and we also have that $\phi(\alpha) - 1/2$ is increasing in α because ϕ is increasing in α on $[1/2, 1)$ by Lemma 7. For any $1/2 < \alpha < \alpha'$, we can thus write

$$\gamma'_\beta(\alpha)\left(\phi(\alpha)-\frac{1}{2}\right) > \gamma'_\beta(\alpha')\left(\phi(\alpha)-\frac{1}{2}\right) \geq \gamma'_\beta(\alpha')\left(\phi(\alpha')-\frac{1}{2}\right),$$

i.e., the first term in the derivative of π is decreasing in α for $\alpha > 1/2$.

For the second term in equation (40), we know that $\gamma_\beta(\alpha)$ is decreasing in α on $(1/2,1]$. We also have that $\phi'(\alpha)$ is decreasing in α on $[1/2,\mu_S/\mu_F]$ because it is concave on this interval by Lemma 7. The derivative $\phi'(\alpha)$ decreases to exactly zero at $\alpha=\mu_S/\mu_F$, and on $[\mu_S/\mu_F,1]$, the derivative $\phi'(\alpha)$ is constant at zero because $\phi(\alpha)$ is constant at 1 on this interval. Therefore, for $1/2<\alpha<\alpha'$, we have

$$\gamma_\beta(\alpha)\phi'(\alpha)\geq\gamma_\beta(\alpha)\phi'(\alpha')\geq\gamma_\beta(\alpha')\phi'(\alpha'),$$

i.e., the second term in equation (40) is also decreasing in α on $(1/2,1]$.

That both terms are decreasing in α (with one strictly decreasing) implies that $\pi'_\beta(\alpha)$ is strictly decreasing in α on $(1/2,1]$. Therefore, we also have that the derivative of $\pi_\beta(\alpha)-\alpha$ is strictly decreasing in α on this interval. By the proof of Lemma 11, we have $\pi_\beta(1/2)-1/2\geq 0$. Since the derivative is strictly decreasing, either $\pi_\beta(\alpha)-\alpha$ is monotonically decreasing on $(1/2,1]$, or it is first increasing and then decreasing (it must have a decreasing interval because $\pi_\beta(1/2)-1/2\geq 0$, while $\pi_\beta(1)-1<0$).

If $\pi_\beta(\alpha)-\alpha$ is monotonically decreasing on $(1/2,1]$, then there can be at most one zero in this interval, and there is at least one zero because the largest equilibrium falls on $(1/2,1]$ by Lemma 11. If it is increasing and then decreasing, then there is no zero in the increasing interval because $\pi_\beta(1/2)-1/2\geq 0$, and there is at most one zero in the decreasing interval by the same reasoning as before. Again, by Lemma 11, there is at least one zero in $(1/2,1]$. Thus, in either case, there is exactly one zero of the function $\pi_\beta(\alpha)-\alpha$ on $(1/2,1]$, which is the only equilibrium in this interval. □

F. Proofs for Section 5

We first state and prove Lemma 13, which provides a closed-form expression for the cumulative system time as a function of the SF fraction α. Then, we give the proof of Proposition 4, that herding is socially optimal.

LEMMA 13 **(Cumulative System Time)**. *Let*

$$\xi_\alpha:=\min\left\{\alpha,\mu_S\Big(\frac{1-\alpha}{\mu_F-\mu_S}\Big)\right\}. \tag{41}$$

The cumulative system time is given by

$$D(\alpha,\mu_S,\mu_F)=\begin{cases}\dfrac{\alpha(1-\alpha)}{\mu_S}+\dfrac{(1-\alpha)^2}{2\mu_S}+\dfrac{\alpha(1-\alpha)}{\mu_F}+\dfrac{\alpha^2}{2\mu_F} & \text{if } \alpha\leq\dfrac{\mu_S}{\mu_F},\\ \dfrac{\alpha(1-\alpha)}{\mu_S}+\dfrac{(1-\alpha)^2}{2\mu_S}+\dfrac{\xi_\alpha(1-\alpha)}{\mu_F}+\dfrac{\alpha^2}{2\mu_S}+\xi_\alpha^2\Big(\dfrac{1}{2\mu_F}-\dfrac{1}{2\mu_S}\Big) & \text{otherwise.}\end{cases} \tag{42}$$

Proof. The overall cumulative system time can be broken down into two components: the cumulative system time for SF customers and the cumulative system time for FS customers. Respectively denoting these by $d_S(\cdot)$ and $d_F(\cdot)$, we have

$$D(\alpha,\mu_S,\mu_F)=d_S(\alpha,\mu_S,\mu_F)+d_F(\alpha,\mu_S,\mu_F). \tag{43}$$

First, we compute the cumulative system time for FS customers. By Lemma 2, for a customer in position y_F, the system time is

$$\mathcal{F}(\alpha;y_F)=\frac{\alpha+y_F}{\mu_S}.$$

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

$$d_F(\alpha,\mu_S,\mu_F)=\int_0^{1-\alpha}\frac{\alpha+y_F}{\mu_S}\mathsf{d}y_F=\frac{\alpha(1-\alpha)}{\mu_S}+\frac{(1-\alpha)^2}{2\mu_S}.$$

For SF customers, by Lemma 1, for a customer in position y_S, the system time is

$$\mathcal{S}(\alpha; y_S) = \begin{cases} \dfrac{y_S + 1 - \alpha}{\mu_F} & y_S \le \mu_S\Big(\dfrac{1-\alpha}{\mu_F - \mu_S}\Big), \\ \dfrac{y_S}{\mu_S} & \text{otherwise.} \end{cases}$$

Recall from equation (41) that

$$\xi_\alpha = \min\left\{\alpha,\, \mu_S\Big(\frac{1-\alpha}{\mu_F - \mu_S}\Big)\right\}.$$

Integrating, we get

$$\begin{aligned} d_S(\alpha, \mu_S, \mu_F) &= \int_0^\alpha \mathcal{S}(\alpha; y_S)\mathsf{d}y_S \\ &= \int_0^{\xi_\alpha} \frac{y_S + 1 - \alpha}{\mu_F}\mathsf{d}y_S + \int_{\xi_\alpha}^{\alpha} \frac{y_S}{\mu_S}\mathsf{d}y_S \\ &= \frac{\xi_\alpha(1-\alpha)}{\mu_F} + \frac{\xi_\alpha^2}{2\mu_F} + \frac{\alpha^2}{2\mu_S} - \frac{\xi_\alpha^2}{2\mu_S}. \end{aligned}$$

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

$$d_S(\alpha, \mu_S, \mu_F) = \begin{cases} \dfrac{\alpha(1-\alpha)}{\mu_F} + \dfrac{\alpha^2}{2\mu_F} & \text{if } \alpha \le \dfrac{\mu_S}{\mu_F}, \\ \dfrac{\xi_\alpha(1-\alpha)}{\mu_F} + \dfrac{\xi_\alpha^2}{2\mu_F} + \dfrac{\alpha^2}{2\mu_S} - \dfrac{\xi_\alpha^2}{2\mu_S} & \text{otherwise.} \end{cases}$$

Substituting our expressions for $d_S(\cdot)$ and $d_F(\cdot)$ into equation (43) gives the formula in equation (42). □

Proof of Proposition 4. First, note that D is continuous in α because the two pieces of equation (42) coincide at the boundary. Differentiating the first piece of equation (42) twice gives

$$\frac{\partial^2 D}{\partial \alpha^2} = -\frac{1}{\mu_S} - \frac{1}{\mu_F} < 0. \tag{44}$$

So, the function is concave in α on $[0, \mu_S/\mu_F]$ and its minimum over this interval will occur at one of the endpoints. We have $D(0, \mu_S, \mu_F) = 1/(2\mu_S)$, while

$$D(\mu_S/\mu_F, \mu_S, \mu_F) = \frac{1}{2\mu_S} + \frac{\mu_S}{2\mu_F^2}\Big(1 - \frac{\mu_S}{\mu_F}\Big) \ge \frac{1}{2\mu_S}. \tag{45}$$

For the case with $\mu_S = \mu_F$, we never have $\alpha > \mu_S/\mu_F$, and the minimizer on $[0, \mu_S/\mu_F]$ is the minimizer on $[0, 1]$. Evaluating the second endpoint, we have $D(1, \mu_S, \mu_F) = 1/(2\mu_S) = D(0, \mu_S, \mu_F)$. So, herding on either route minimizes the cumulative system time over $[0, 1]$.

For the rest of the proof, we assume that $\mu_S < \mu_F$. In this case, on $[0, \mu_S/\mu_F]$, the function has a unique minimizer of 0 by equation (45), with a cumulative system time of $1/(2\mu_S)$. We must also consider the second piece of the function that governs on $[\mu_S/\mu_F, 1]$. Differentiating the second piece of equation (42), we have

$$\frac{\partial D}{\partial \alpha} = -\frac{\mu_S(1-\alpha)}{\mu_F(\mu_F - \mu_S)} \le 0 \text{ for } \alpha \le 1. \tag{46}$$

Therefore, the cumulative system time is decreasing for $\alpha \in [\mu_S/\mu_F, 1]$, and it is thus minimized at $\alpha = 1$ on this interval. Substituting $\alpha = 1$ into the second piece of the function gives $D(1, \mu_S, \mu_F) = 1/(2\mu_S) = D(0, \mu_S, \mu_F)$. Thus, the cumulative system time is again minimized at $\alpha = 0$ or 1, i.e., when customers herd on the same route. □

G. Extension: Randomization with Unequal Probabilities Under Missing Anecdotes

We now prove that our main finding in Theorem 1—that high discernibility produces herding—continues to hold when we relax our assumption from the base model that a customer missing anecdotes from one route randomizes with equal probability among the two routes.

In this section, we assume that a customer missing anecdotes from one route chooses the route that is present in her sample with probability $1/2 < \kappa < 1$ and the other route with probability $1-\kappa$. This assumption reflects a customer who favors the route that is present in her sample. Since receiving all anecdotes from the more-prevalent route is more likely than receiving all anecdotes from the less-prevalent route, favoring the route that is present in the sample could suggest that the customer has some intuition that "following the crowd" is preferred.

With the following result, we establish that high discernibility continues to produce herding under this relaxed assumption. Recall that $\alpha_{K,\beta}$ denotes the largest equilibrium SF fraction from the base model with equal-probability randomization under missing anecdotes, and now denote by $\alpha_{K,\beta,\kappa}$ the largest equilibrium SF fraction under the relaxed assumption with probability κ.

PROPOSITION 5 **(Convergence to Herding with Unequal Randomization)**. *The largest equilibrium $\alpha_{K,\beta,\kappa}$ converges to herding on route SF as β increases, i.e., we have*

$$\lim_{\beta\to\infty} \alpha_{K,\beta,\kappa} = 1.$$

Furthermore, if $\alpha_{K,\beta} > 1/2$, we have $\alpha_{K,\beta,\kappa} > \alpha_{K,\beta}$.

The proof requires a lemma. Denote by $\mathcal{O}_{SF}$ ($\mathcal{O}_{FS}$)—$\mathcal{O}$ for Only—the event that all of a customer's anecdotes come from route SF (FS).

LEMMA 14 **(More Likely to Get All Anecdotes from Most Prevalent Route)**. *For $\alpha > 1/2$, we have*

$$\mathsf{Pr}[\mathcal{O}_{SF}|\mathcal{O}_{SF}\cup\mathcal{O}_{FS}] > \frac{1}{2} > \mathsf{Pr}[\mathcal{O}_{FS}|\mathcal{O}_{SF}\cup\mathcal{O}_{FS}].$$

Proof of Lemma. Define

$$h_\beta(x) := \frac{\beta x}{\beta x + (1-x)}.$$

By equations (1) and (2) and following the sampling process described in Section 3.1, we have

$$\mathsf{Pr}[\mathcal{O}_{SF}] = \alpha\left(1-h_\beta\left(1-\alpha\right)\right)^{K-1} \quad \text{and} \quad \mathsf{Pr}[\mathcal{O}_{FS}] = (1-\alpha)\left(1-h_\beta\left(\alpha\right)\right)^{K-1}. \tag{47}$$

Because $\beta \geq 1$, the function $h_\beta(x)$ is increasing in x. Then, since $\alpha > 1/2$ implies $\alpha > 1-\alpha$, we have $h_\beta(\alpha) > h_\beta(1-\alpha)$ and thus $1-h_\beta(1-\alpha) > 1-h_\beta(\alpha)$. Combined with equation (47) and again using $\alpha > 1-\alpha$, this implies

$$\mathsf{Pr}[\mathcal{O}_{SF}] = \alpha\left(1-h_\beta\left(1-\alpha\right)\right)^{K-1} > (1-\alpha)\left(1-h_\beta\left(\alpha\right)\right)^{K-1} = \mathsf{Pr}[\mathcal{O}_{FS}],$$

which in turn gives

$$\mathsf{Pr}[\mathcal{O}_{SF}|\mathcal{O}_{SF}\cup\mathcal{O}_{FS}] = \frac{\mathsf{Pr}[\mathcal{O}_{SF}]}{\mathsf{Pr}[\mathcal{O}_{SF}]+\mathsf{Pr}[\mathcal{O}_{FS}]} > \frac{1}{2}. \tag{48}$$

We also have $\mathsf{Pr}[\mathcal{O}_{FS}|\mathcal{O}_{SF}\cup\mathcal{O}_{FS}] = 1-\mathsf{Pr}[\mathcal{O}_{SF}|\mathcal{O}_{SF}\cup\mathcal{O}_{FS}] < 1/2$, and the result follows.

As an aside, we remark that the reverse result holds for $\alpha < 1/2$ by symmetry, i.e., for $\alpha < 1/2$ we have $\mathsf{Pr}[\mathcal{O}_{FS}] > \mathsf{Pr}[\mathcal{O}_{SF}]$ and $\mathsf{Pr}[\mathcal{O}_{FS}|\mathcal{O}_{SF}\cup\mathcal{O}_{FS}] > 1/2 > \mathsf{Pr}[\mathcal{O}_{SF}|\mathcal{O}_{SF}\cup\mathcal{O}_{FS}]$. □

Proof of Proposition 5. Recall that $\mathcal{B}$ is the event that a customer receives at least one anecdote from each route, and observe that $\mathcal{B}^{\mathrm{C}} = \mathcal{O}_{SF} \cup \mathcal{O}_{FS}$. The more general version of the response function (3) without the assumption of equal-probability randomization under missing anecdotes is then given by

$$\pi_{K,\beta,\cdot}(\alpha) = \mathsf{Pr}[\bar{s}_{K,\beta}(\alpha) \leq \bar{f}_{K,\beta}(\alpha)|\mathcal{B}] \cdot \gamma_{K,\beta}(\alpha) + \mathsf{Pr}[\text{Customer chooses } SF|\mathcal{O}_{SF} \cup \mathcal{O}_{FS}] \cdot (1 - \gamma_{K,\beta}(\alpha)). \quad (49)$$

Now, define $w(x) := \kappa x + (1-\kappa)(1-x)$, and let $\delta_\alpha = \mathsf{Pr}[\mathcal{O}_{SF}|\mathcal{O}_{SF} \cup \mathcal{O}_{FS}]$ for a given SF fraction α (note that the sampling process itself is the same regardless of the routing decision rule for customers with missing anecdotes). Under the assumption that a customer who receives anecdotes of only one route chooses that route with probability κ, we have

$$\mathsf{Pr}[\text{Customer chooses } SF|\mathcal{O}_{SF} \cup \mathcal{O}_{FS}] = \mathsf{Pr}[\mathcal{O}_{SF}|\mathcal{O}_{SF} \cup \mathcal{O}_{FS}]\kappa + \mathsf{Pr}[\mathcal{O}_{FS}|\mathcal{O}_{SF} \cup \mathcal{O}_{FS}](1-\kappa) = w(\delta_\alpha). \quad (50)$$

Substituting the RHS of equation (50) and including κ in the notation, the response function (49) becomes

$$\pi_{K,\beta,\kappa}(\alpha) = \mathsf{Pr}[\bar{s}_{K,\beta}(\alpha) \leq \bar{f}_{K,\beta}(\alpha)|\mathcal{B}] \cdot \gamma_{K,\beta}(\alpha) + w(\delta_\alpha)(1 - \gamma_{K,\beta}(\alpha)). \quad (51)$$

Straightforward arguments building on Lemmas 3 and 4 imply that $\pi_{K,\beta,\kappa}(\alpha)$ in equation (51) is continuous in α over $[0,1]$ with appropriate boundary conditions $\pi_{K,\beta,\kappa}(0) = 1-\kappa$ and $\pi_{K,\beta,\kappa}(1) = \kappa$, and that the set of equilibria, i.e., $\{\alpha : \pi_{K,\beta,\kappa}(\alpha) = \alpha\}$ is non-empty and compact; we omit these arguments for brevity. A largest equilibrium therefore exists, which we denote by $\alpha_{K,\beta,\kappa}$.

Next, observe that (i) $w(1/2) = 1/2$ and (ii) since $\kappa > 1/2$, the function $w(x)$ is increasing in x because $w'(x) = 2\kappa - 1 > 0$. For $\alpha > 1/2$, by Lemma 14, we have $\delta_\alpha > 1/2$, and since $w(x)$ is increasing, this implies $w(\delta_\alpha) > w(1/2) = 1/2$. Combining this with equation (51) gives

$$\begin{aligned} \pi_{K,\beta,\kappa}(\alpha) &= \mathsf{Pr}[\bar{s}_{K,\beta}(\alpha) \leq \bar{f}_{K,\beta}(\alpha)|\mathcal{B}] \cdot \gamma_{K,\beta}(\alpha) + w(\delta_\alpha)(1 - \gamma_{K,\beta}(\alpha)) \\ &> \mathsf{Pr}[\bar{s}_{K,\beta}(\alpha) \leq \bar{f}_{K,\beta}(\alpha)|\mathcal{B}] \cdot \gamma_{K,\beta}(\alpha) + \frac{1}{2}(1 - \gamma_{K,\beta}(\alpha)) \\ &= \pi_{K,\beta}(\alpha), \end{aligned} \quad (52)$$

where the last equality holds by equation (3). By definition, we have $\pi_{K,\beta}(\alpha_{K,\beta}) = \alpha_{K,\beta}$. Also, by assumption, we have β large enough that $\alpha_{K,\beta} > 1/2$ (such β exists because $\lim_{\beta\to\infty} \alpha_{K,\beta} = 1$ by Theorem 1). By equation (52), we then have $\pi_{K,\beta,\kappa}(\alpha_{K,\beta}) > \pi_{K,\beta}(\alpha_{K,\beta}) = \alpha_{K,\beta}$. Because $\pi_{K,\beta,\kappa}(1) = \kappa < 1$, there must be an equilibrium on the interval $(\alpha_{K,\beta}, 1]$ by the intermediate value theorem, so we have $\alpha_{K,\beta,\kappa} > \alpha_{K,\beta}$, as desired.

Finally, since $\alpha_{K,\beta} < \alpha_{K,\beta,\kappa} \leq 1$ for all β sufficiently large and also $\lim_{\beta\to\infty} \alpha_{K,\beta} = 1$ by Theorem 1, we must have $\lim_{\beta\to\infty} \alpha_{K,\beta} = 1$ by the sandwich theorem. This completes the proof. □