M&SOM MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Vol. 6, No. 1, Winter 2004, pp. 23–40
ISSN 1523-4614 | EISSN 1526-5498 | 04 | 0601 | 0023

informs

DOI 10.1287/msom.1030.0030

Strategically Seeking Service: How Competition Can Generate Poisson Arrivals

Martin A. Lariviere, Jan A. Van Mieghem
Kellogg School, Northwestern University, 2001 Sheridan Road, Evanston, Illinois 60208-2009
{m-lariviere@kellogg.northwestern.edu, vanmieghem@kellogg.northwestern.edu}

We consider a simple game in which strategic agents select arrival times to a service facility. Agents find congestion costly and, hence, try to arrive when the system is underutilized. Working in discrete time, we characterize pure-strategy Nash equilibria for the case of ample service capacity. In this case, agents try to spread themselves out as much as possible and their self-interested actions will lead to a socially optimal outcome if all agents have the same well-behaved delay cost function. For even modest sized problems, the set of possible pure-strategy Nash equilibria is quite large, making implementation potentially cumbersome. We consequently examine mixed-strategy Nash equilibria and show that there is a unique symmetric Nash equilibrium. Not only is this equilibrium independent of the number of agents and their individual delay cost functions, the arrival pattern it generates approaches a discrete-time Poisson process as the number of agents and arrival points gets large. Our results extend to the case of time varying preferences. With an appropriate initialization, the results also extend to a system with limited capacity. Our model lends support to the traditional literature on managing service systems. This work has generally ignored customers strategically choosing arrival times. Rather it is commonly assumed that customers seek service according to some well-behaved process (e.g., that interarrival times follow a renewal process). We show that assuming Poisson arrivals is an acceptable assumption even with strategic customers if the population is large and the horizon is long.

Key words: Poisson process; service management; queuing; game theory; mixed-strategy equilibrium
History: Received: January 21, 2003; days with authors: 111; revisions: 2; average review cycle time: 67 days; Senior Editor: Noah Gans.

1. Introduction

Woody Allen has claimed that 80% of success is showing up, but there are instances in which Allen's dictum seems an underestimate. Consider errands such as dropping off dry cleaning or mailing a package. For such mundane services, convenience is often as important as price. Success then is almost entirely about showing up. If one arrives when the service facility is lightly loaded, one encounters little or no delay. If one arrives when the facility is heavily utilized, waits are likely significant. Minimizing the cost of the errand consequently depends on when one seeks service relative to when others arrive at the facility.

Here, we study how strategic agents should seek service. A set of agents patronizes a service facility. All are delay sensitive and choose arrival times from a discrete set simultaneously. A self-interested agent maximizes her net utility (i.e., the difference between her value of the service and the congestion cost she incurs) by arriving with as few other agents as possible. Each agent consequently takes into account the actions of the others when deciding when to show up. We first characterize possible Nash equilibria. In particular, we examine how numerous and complicated equilibria are. Implementing a Nash equilibrium requires that agents "coordinate" their actions either explicitly or implicitly: Each agent must know which Nash equilibrium is being played (assuming there is more than one) and what action she is supposed to take under the chosen equilibrium. Such coordination may be difficult if the set of Nash equilibria is large or the chosen equilibrium is complicated.

We next consider whether any of the possible Nash equilibria can be linked to standard arrival processes assumed in the literature. We are most interested in this point. The management of service facilities is generally predicated on customers arriving according to

a well-understood process. For example, one might assume that the time between arrivals is given by a renewal process. But is there any reason to believe that customers strategically trying to minimize their individual delay costs should be so obliging? The arrival process generated by strategic customers may bear no relation to standard assumptions, calling into question the recommendations of models built on customers arriving according to a convenient process. Fortunately, we have an encouraging result: In our model, there exists a simple, plausible equilibrium in which competing customers generate Poisson arrivals. Thus, customers arrive according to a commonly assumed stochastic process as a *consequence* of strategic interaction.

Below, we show that many pure-strategy Nash equilibria exist. These equilibria are all independent of the agents' delay cost functions. In any pure-strategy Nash equilibrium, customers spread out as much as possible; the number of arrivals to any two periods differs by at most one. For symmetric problems, pure-strategy Nash equilibria minimize total system delay costs subject to a mild regularity condition. Self-interested agents thus implement socially efficient outcomes.

While a pure-strategy Nash equilibrium is attractive, its implementation is cumbersome. All agents must agree on the specific equilibrium being played. Such coordination is possible if agents choose sequentially and choices are observable, i.e., an appointment system results in a Stackelberg equilibrium which is identical to some pure-strategy Nash equilibrium. However, such an arrangement may not be possible. We therefore consider mixed-strategy Nash equilibria. Again, there is a large number of outcomes, but equilibria may now depend on agents' delay cost functions. We show that there is a unique symmetric equilibrium. Further, this equilibrium is independent of the number of agents and their individual delay cost functions. In this equilibrium, each agent puts equal probability on every arrival time and, thus, the number of arrivals in any time period has a binomial distribution. The distribution of arrivals per time period is stationary over the horizon and converges to a Poisson distribution as the number of agents and time periods gets large. In the limit, the distribution of arrivals across periods is independent. Hence, the arrival pattern converges to a discrete-time Poisson process as the number of agents and time periods gets large.

In §2, we provide a brief literature review. In §3, we present the model and develop results assuming that agents have preferences independent of when they receive service and that agents must be served in the period in which they arrive. In §4, we consider time-varying preferences. In §5, we suppose that agents may have to wait one or more periods to be served. §6 concludes.

2. Related Literature

Much of economics is based on the observation that, all else being equal, people prefer the same goods at a lower cost. As most find waiting inconvenient, a natural generalization is to assume that people prefer to avoid congestion and the concomitant delay. How rational agents respond to expected delays is consequently an active area of study. Customers have been assumed to be sophisticated in whether they join or balk from a queue (Naor 1969, Yechiali 1971), in how they submit work (Dewan and Mendelson 1990, Stidham 1992), and in how they declare their priority class (Mendelson 1985, Mendelson and Whang 1990, Van Mieghem 2000). However, all of this work assumes that customers arrive according to a renewal process. Hassin and Haviv (2003) provide an excellent review of this literature.

Economists have considered games in which payoffs depend on the order of arrivals or departures. For example, in a small market, multiple firms may not be able to cover fixed costs although one firm alone could be profitable. A war of attrition results, and the analysis focuses on when firms exit the market. See Fudenberg and Tirole (1986). Alternatively, a firm might benefit from being the first to enter a market. Fudenberg and Tirole (1985) present an example of such a preemption game. See Park and Smith (2003) for a general formulation of these rank-order timing games and additional references. In our model, agents care only about the congestion they encounter, not the order in which they arrive.

In Vickrey (1969), commuters use a bottleneck stretch of road. A queue forms when the arrival rate is above the road's capacity. He examines the arrival rate to the bottleneck and the optimality of toll

schemes. His commuters balance the cost of leaving early or arriving late but are not sensitive to congestion while using the road. In our model, an agent's costs increase with congestion. All agents prefer to be the only one at the service facility.

In Ostrovsky and Schwarz (2003), all agents must arrive for processing to start, but all find waiting for the last arrival costly. Coordination requires simultaneous arrivals, but independent agents may have an incentive to be late. In our model, coordination requires agents to spread out their arrival times and failure occurs when too many arrive at once.

Most relevant to our work is Glazer and Hassin (1983). A random number of symmetric customers with linear delay costs seek service over some horizon. They may queue before the service facility opens. Service times are exponentially distributed. Working in continuous time, they derive equations that a symmetric Nash equilibrium must satisfy. The arrival rate is generally not constant over the horizon. After an early spike, expected arrivals taper off. The equilibrium depends on system parameters. We work in discrete time which allows us to ignore details of the service process and accommodate asymmetric agents. Also, Glazer and Hassin (1983) focus on transient behavior over a fixed time horizon. We are interested in limiting (stationary) results as the horizon and number of customers gets large.

Rapoport et al. (2003) and Seale et al. (2003) calculate the symmetric mixed-strategy Nash equilibrium for a particular discrete-time version of Glazer and Hassin (1983) and compare the theoretical equilibrium with how experimental subjects actually choose arrival times. They find support for mixed-strategy play at an aggregate level. In our model, the unique symmetric equilibrium involves mixed strategies.

3. Model Basics and Equilibria with Ample Capacity

We consider a finite horizon divided into $T \geq 2$ equal-sized time periods or "bins" (we will use the terms interchangeably). Without loss of generality, fix the length of a time period at one. There are $M \geq 2$ agents or customers who seek service over the horizon by choosing an arrival bin. Let $\alpha_t = 0, \ldots, M$ be the number of customers arriving to bin t, and $A(t) = \sum_{i=1}^{t} \alpha_i$ be the associated cumulative arrival process. Let $\lambda = M/T$ be the average number of customers per bin. We adopt the convention that all customers arrive at the start of the time period. For now we assume that the system has ample capacity, i.e., there is sufficient resources to serve all M customers in one time period. A customer arriving in time period t is therefore certain to receive service in that time period. No customers remain in the system from period t to period $t+1$, so the only customers in the system during period t are the α_t who arrive in that period. We relax this assumption in §5.

Customer m values service at $V_m > 0$ regardless of the period in which she is served. All customers prefer to avoid congestion. Let $W_m(\alpha_t)$ denote agent m's expected disutility of being one of α_t arrivals in period t. (α_t includes agent m.) Agent m's objective is to maximize her net utility $U_m(\alpha_t) = V_m - W_m(\alpha_t)$ through her choice of arrival bin t. Equivalently, agent m seeks to minimize her expected congestion or delay cost $W_m(\alpha_t)$.

This formulation embeds an important assumption. V_m is independent of t, so agent m's net utility depends on the time bin only through the number of arrivals α_t. Thus, we are assuming a nonurgent service for which agents care only about the wait they encounter and not about the exact timing of service. We consider time-varying preferences in §4.

We allow for significant flexibility in modeling the expected delay cost $W_m(\alpha_t)$; all we require is that $W_m(\alpha_t)$ is strictly increasing in α_t, the number of arrivals to bin t. We will emphasize two basic forms of W_m depending on whether the agents are served sequentially or in a batch. Thus, W_m captures both agent preferences and the physics of how the service operates. In the sequential-service case, agent m incurs a cost $C_m(k) \geq 0$ for $k = 1, \ldots, \alpha_t$ if she is the kth customer to be served for some function C_m. On arrival, the agents are randomly ordered such that each agent has an equal probability of being in any position in line. Given $\alpha_t \geq 1$, agent m's expected delay costs are

$$W_m(\alpha_t) = \frac{1}{\alpha_t} \sum_{k=1}^{\alpha_t} C_m(k). \tag{1}$$

We assume that $C_m(k)$ increases sufficiently fast so that $W_m(\alpha_t)$ is strictly increasing.[1]

[1] This requires that $C_m(\alpha+1) > \alpha^{-1} \sum_{k=1}^{\alpha} C_m(k)$ for $\alpha = 1, \ldots, M-1$.

On occasion, we will consider two special cases of (1). First, suppose that delay costs are linear. $C_m(k) = \theta_m \times (k-1)$ for $\theta_m > 0$. We then have

$$W_m(\alpha_t) = \frac{\theta_m}{\alpha_t}\sum_{k=1}^{\alpha_t}(k-1) = \theta_m\frac{\alpha_t - 1}{2}. \tag{2}$$

Alternatively, delay costs can increase at an increasing rate. Suppose $C_m(k) = \theta_m^{k-1} - 1$ for $\theta_m > 1$. The resulting expected delay costs are

$$W_m(\alpha_t) = \frac{1}{\alpha_t}\sum_{k=1}^{\alpha_t}\theta_m^{k-1} - 1 = \frac{\theta_m^{\alpha_t} - 1}{\alpha_t(\theta_m - 1)} - 1. \tag{3}$$

In a batch setting, all arrivals are served simultaneously but the quality of service falls with the number of agents being served

$$W_m(\alpha_t) = \theta_m - \frac{\theta_m}{\alpha_t} \quad \text{for } \theta_m > 0. \tag{4}$$

If m's delay cost is (2), the agent only cares about the average number arrivals to bin t and, thus, behaves as if she were risk neutral. In contrast, (3) is convex, making the agent's net utility concave in arrivals.[2] Hence, such an agent behaves as if she were risk averse and prefers to avoid variability. The reverse holds in a batch setting. The agent's delay cost is concave and her net utility is convex. With batch costs, an agent behaves as if she were risk seeking and prefers variability in the number of arrivals. We exploit these properties below.

Agents choose arrival times simultaneously and irrevocably. Agent m cannot balk or condition her choice on the decisions of others or on realized queue lengths. For convenience, we assume agent m knows the number of agents M, the number of arrival bins T, and the delay cost functions of all agents. Below we note instances in which common knowledge of all system parameters is not required.

3.1. Pure-Strategy Nash Equilibria

Agent m would like to seek service in the bin that would maximize her expected net benefit, $U_m(\alpha_t)$. Of course, that benefit depends on the actions of the other agents. We must look for equilibrium arrival patterns. For now we restrict our attention to pure-strategy Nash equilibria in which each agent reports deterministically to one bin. Defining such a Nash equilibrium requires some notation. Let $\mathbf{e}_j$ denote the jth $1 \times T$ unit vector, and let $\boldsymbol{\pi}_m$ denote agent m's strategy: $\boldsymbol{\pi}_m = \mathbf{e}_t$ if agent m reports to bin t. Let $\boldsymbol{\Pi} = (\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_M)$ be a strategy profile for the M agents, and let $\boldsymbol{\Pi}_{-m} = (\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_{m-1}, \boldsymbol{\pi}_{m+1}, \ldots, \boldsymbol{\pi}_M)$. Define $\boldsymbol{\alpha}(\boldsymbol{\Pi})$ as the arrival vector that results from $\boldsymbol{\Pi}$:

$$\boldsymbol{\alpha}(\boldsymbol{\Pi}) = (\alpha_1(\boldsymbol{\Pi}), \ldots, \alpha_T(\boldsymbol{\Pi})) = \sum_{j=1}^{M}\boldsymbol{\pi}_j$$

and $\alpha(\boldsymbol{\Pi}_{-m})$ as the arrival vector of all agents but m:

$$\alpha(\boldsymbol{\Pi}_{-m}) = (\alpha_1(\boldsymbol{\Pi}_{-m}), \ldots, \alpha_T(\boldsymbol{\Pi}_{-m})) = \sum_{j\neq m}^{M}\boldsymbol{\pi}_j.$$

Finally, let $B_m(\boldsymbol{\Pi})$ denote the bin chosen by agent m under strategy profile $\boldsymbol{\Pi}$. $B_m(\boldsymbol{\Pi}) = t$ if $\boldsymbol{\pi}_m = \mathbf{e}_t$. We can now state that $\boldsymbol{\Pi}^*$ is a pure-strategy Nash equilibrium if the following holds for $m = 1, \ldots, M$:

$$V_m - W_m(\alpha_{B_m(\boldsymbol{\Pi}^*)}(\boldsymbol{\Pi}^*)) \geq V_m - W_m(\alpha_t(\boldsymbol{\Pi}^*_{-m}) + 1) \quad \text{for } t = 1, \ldots, T.$$

In words, a pure-strategy Nash equilibrium requires that holding the proposed actions of others fixed, an agent cannot improve her payoff by unilaterally moving to a different bin.

THEOREM 1. *A strategy profile $\boldsymbol{\Pi}^*$ is a pure-strategy Nash equilibrium if and only if $\alpha_t(\boldsymbol{\Pi}^*) - \alpha_s(\boldsymbol{\Pi}^*) \leq 1$ for all $s, t = 1, \ldots, T$.*

PROOF. Suppose that $\boldsymbol{\Pi}^*$ is a Nash equilibrium and that $B_m(\boldsymbol{\Pi}^*) = t$. Consider agent m's incentive to deviate. As W_m is strictly increasing, she has no interest in moving to any bin s such that $\alpha_s(\boldsymbol{\Pi}^*) \geq \alpha_t(\boldsymbol{\Pi}^*)$. Suppose there is a bin s' such that $\alpha_t(\boldsymbol{\Pi}^*) > \alpha_{s'}(\boldsymbol{\Pi}^*)$. Deviating is profitable if $\alpha_t(\boldsymbol{\Pi}^*) > \alpha_{s'}(\boldsymbol{\Pi}^*_{-m}) + 1 = \alpha_{s'}(\boldsymbol{\Pi}^*) + 1$. Thus, if $\boldsymbol{\Pi}^*$ is an equilibrium, it must be the case that $\alpha_t(\boldsymbol{\Pi}^*) - \alpha_s(\boldsymbol{\Pi}^*) \leq 1$ for all $s, t = 1, \ldots, T$.

Now suppose $\alpha_t(\boldsymbol{\Pi}^*) - \alpha_s(\boldsymbol{\Pi}^*) \leq 1$ for all $s, t = 1, \ldots, T$ and $B_m(\boldsymbol{\Pi}^*) = t$. If she were to move to bin s, the number of arrivals to s would be $\alpha_s(\boldsymbol{\Pi}^*_{-m}) + 1 = \alpha_s(\boldsymbol{\Pi}^*) + 1 \geq \alpha_t(\boldsymbol{\Pi}^*)$. Hence, she has no reason to unilaterally deviate from t, and $\boldsymbol{\Pi}^*$ is an equilibrium. □

[2] Convexity (concavity) is overly restrictive. W_m is defined for natural numbers and need not be continuous. We use convexity (concavity) as short hand for increasing (decreasing) first differences, i.e., $W_m(\alpha+1) - W_m(\alpha) \geq [\leq] W_m(\alpha) - W_m(\alpha-1)$.

In a pure-strategy Nash equilibrium each bin must have a "Yogi-Berra" property: In equilibrium, no one goes there anymore; it's too crowded. Each agent is content to report to her assigned bin because, holding everyone else's decision constant, she cannot find one sufficiently less congested. Compared to her current assignment every other bin is either as crowded, more crowded, or will become as crowded if she were to deviate by moving to it. Much as drops of water in a glass settle at a uniform level to minimize their potential energy, agents here spread out as close to uniformly as possible to minimize their expected delay. Deviations from a uniform level occur only because agents are not infinitely divisible.

Theorem 1 requires only that delay cost functions are strictly increasing. Although it is convenient to assume delay cost functions are commonly known, it is unnecessary when implementing a pure-strategy Nash equilibrium. A pure-strategy Nash equilibrium for a given set of strictly increasing delay cost functions is an equilibrium for any set of strictly increasing delay cost functions.

To construct a pure-strategy Nash equilibrium, let $\lfloor x \rfloor$ denote the smallest integer less than or equal to x. Let $\underline{\lambda} = \lfloor \lambda \rfloor$ and $\tau = M - \underline{\lambda}T$. Clearly, $\tau = 0$ if $\lambda = \underline{\lambda}$ and $0 < \tau < T$ if $\lambda > \underline{\lambda}$. If one assigns $\underline{\lambda}T$ customers such that each bin has $\underline{\lambda}$ agents, τ agents remain. A Nash equilibrium $\mathbf{\Pi}_1^*$ can be formed by assigning each of these τ agents to a distinct bin. As all agents are assigned an arrival time and each agent is assigned to only one bin, the vector is feasible. As the maximum difference in arrivals is one, $\mathbf{\Pi}_1^*$ is a pure-strategy Nash equilibrium. An arbitrary pure-strategy Nash equilibrium thus has $\tau \geq 0$ "crowded" bins with $\underline{\lambda}+1$ arrivals and $T - \tau > 0$ "uncrowded" bins with only $\underline{\lambda}$ arrivals, i.e., there is at least one bin with just $\underline{\lambda}$ arrivals.

We now consider how many Nash equilbria exist and whether Nash equilibria can be socially efficient in the sense of minimizing total delay costs. To speak to the first point, it is easy to see that because $T \geq 2$ and $M \geq 2$, a Nash equilibrium $\mathbf{\Pi}_1^*$ is not unique. If one merely interchanges an agent assigned to bin t with an agent assigned to bin $s \neq t$, one has created a new equilibrium $\mathbf{\Pi}_2^*$. Indeed, the set of pure-strategy Nash equilibria can be quite large.

Corollary 1. *Let $\bar{\Pi}$ denote the set of all possible pure-strategy Nash equilibria. Let $\|X\|$ denote the cardinality of the set X.*

$$\|\bar{\Pi}\| = \frac{M!}{\underline{\lambda}!^{(T-\tau)}(\underline{\lambda}+1)!^{\tau}}\binom{T}{\tau}.$$

Proof. See the Appendix. □

Table 1 gives the number of pure-strategy Nash equilibria for sample values of M and T. For even modest-sized problems, there are thousands—if not millions—of equilibria to choose from.

To say something about the social optimality of a pure-strategy Nash equilibrium, we must consider symmetric agents. Suppose all agents have the same delay cost function $W(\alpha_t)$. If α_t agents are assigned to bin t, the resulting social congestion costs for that bin are $S(\alpha_t) = \alpha_t W(\alpha_t)$. If $S(\alpha_t)$ is well behaved, pure-strategy Nash equilibria are socially efficient in the sense of minimizing total delay costs.

Theorem 2. *If $S(\alpha_t)$ is weakly convex, the minimum social cost is $\bar{S} = \tau S(\underline{\lambda}+1) + (T-\tau)S(\underline{\lambda})$ and any pure-strategy Nash equilibrium $\mathbf{\Pi}^*$ achieves cost $\bar{S}$. If $S(\alpha_t)$ is strictly convex, only implementing a pure-strategy Nash equilibrium can achieve cost $\bar{S}$.*

Proof. Because any pure-strategy Nash equilibrium must have $\tau \geq 0$ bins with $\underline{\lambda}+1$ arrivals and $T - \tau > 0$ bins with only $\underline{\lambda}$ arrivals, all pure-strategy equilibria must have social cost $\bar{S}$. Now suppose there exists an arrival vector $\alpha' = \{\alpha'_1, \ldots, \alpha'_T\}$ that minimizes costs but that cannot be implemented by a pure-strategy Nash equilibrium (i.e., $\alpha(\mathbf{\Pi}^*) \neq \alpha'$ for all $\mathbf{\Pi}^* \in \bar{\Pi}$). Because α' is not the result of a pure-strategy Nash equilibrium, it must be that $\alpha'_t > \alpha'_s + 1$

Table 1 How the Number of Equlibria Varies with M and T

M	T	Number of Pure-Strategy Nash Equilibria
5	10	30,240
10	10	3,628,800
15	10	1.03E+13
20	10	2.38E+15
10	20	6.70E+11
20	20	2.43E+18
30	20	4.79E+34
40	20	7.78E+41

for some s and t. However, because $S(\alpha_t)$ is weakly convex,

$$S(\alpha'_t) + S(\alpha'_s) \geq S(\alpha'_t - 1) + S(\alpha'_s + 1). \tag{5}$$

Thus, costs are weakly lowered by transferring agents from t to s until $\alpha'_t - \alpha'_s \leq 1$ for all s and t. Such an arrival vector could then be implemented by a pure-strategy Nash equilibrium. Note that if (5) holds as a strict inequality for all α'_t and α'_s, minimum cost can only be achieved by a pure-strategy Nash equilibrium. □

Just as no agent can single-handedly lower her congestion cost under a pure-strategy Nash equilibrium, there is no way a central planner can lower system costs in a symmetric problem. Self-interested agents will choose a socially efficient outcome that minimizes total waiting costs. The required condition is fairly weak. It is obviously satisfied by any weakly convex delay cost function. Some concave cost functions such as (4) also satisfy it.

Pure-strategy Nash equilibria thus have some attractive properties, but with such a large number of equilibria, implementation is potentially an issue. The following offers a possible means for selecting one.

THEOREM 3. *Suppose agents select bins sequentially. Agent m observes all prior arrivals before choosing a bin. If agent m is indifferent among bins, assume she chooses randomly among bins with equal probability. This procedure leads to an element of* $\bar{\Pi}$.

PROOF. For the last agent, there must be at least one bin with $\underline{\lambda}$ or fewer agents. She prefers this to any bin with $\underline{\lambda}+1$ or more agents. Similarly, no earlier agent will choose a bin already occupied by $\underline{\lambda}+1$ agents. Thus, the final arrangement of agents will have bins with either $\underline{\lambda}$ or $\underline{\lambda}+1$ agents and must be a pure-strategy Nash equilibrium by Theorem 1. □

Customers benefit from an appointment system that enforces sequential choice with visibility. An appointment scheme results in a Stackelberg game, the outcome of which is also a pure-strategy Nash equilibrium. The agents again do not need to know everyone's delay cost function, but they do need to know the total number of agents to know when a bin is full.[3] In addition, some assumption about how indifferent agents choose is required. If the initial sequencing of agents is random, the assumption that indifferent agents randomize implies that any element of $\bar{\Pi}$ is a feasible outcome. This does not hold for other behavioral assumptions. Suppose instead that an agent choosing among bins having $\underline{\lambda}$ occupants always selects the earliest bin. Then, if $M > T$, the first $\underline{\lambda}(T-\tau)$ agents always opt to arrive in the last $(T-\tau)$ time periods because they anticipate that the final τ agents will choose to arrive in one of the first τ periods. Hence, the first τ bins will always be the crowded bins with $\underline{\lambda}+1$ occupants.

3.2. Mixed-Strategy Nash Equilibria

Sequential choice requires less upfront coordination and information than picking an arbitrary pure-strategy Nash equilibrium, but it may not always be feasible. Coordination necessitates sequencing the agents and taking appointments. In addition, the agents must know the total number of customers attempting to use the system. Consequently, a mixed-strategy Nash equilibrium may be a more plausible outcome. Here agents randomize over their possible actions, i.e., which bin to select, so an agent's strategy is a probability distribution over the set of bins. We now have $\boldsymbol{\pi}_m = (\pi_m^1, \ldots, \pi_m^T)$ for $\pi_m^t \geq 0$ and $\sum_{t=1}^T \pi_m^t = 1$. Note that the pure strategies considered above are subsumed in this formulation by considering degenerate distributions (i.e., by allowing $\boldsymbol{\pi}_m$ to be a unit vector). The set of mixed-strategy Nash equilibria is thus at least as large as the set of pure-strategy Nash equilibria.

The number of arrivals to bin t, $\alpha_t(\boldsymbol{\Pi})$, is now a random variable that depends on the strategy profile $\boldsymbol{\Pi}$. Let $\alpha_t(\boldsymbol{\Pi}_{-m})$ denote the number of arrivals to t excluding agent m. $\alpha_t(\boldsymbol{\Pi})$ takes values from zero to M while $\alpha_t(\boldsymbol{\Pi}_{-m})$ takes values from zero to $M-1$. If agent m's realized bin choice is t, her expected delay costs are $\mathbb{E}[W_m(\alpha_t(\boldsymbol{\Pi}_{-m})+1)]$. We can therefore define $\mathscr{W}_m(\boldsymbol{\Pi})$, m's expected delay cost when all agents follow mixed strategies $\boldsymbol{\Pi}$, as

$$\mathscr{W}_m(\boldsymbol{\Pi}) = \sum_{t=1}^{T} \pi_m^t \mathbb{E}[W_m(\alpha_t(\boldsymbol{\Pi}_{-m})+1)].$$

[3] We are assuming that agents are signing up on an open list and are determining for themselves when a bin is too crowded to join. Alternatively, a central coordinator can control the appointment book. Now only the coordinator would need to know the total number of agents, but each agent must be told which bins are still open when she chooses.

We say that a random variable X is stochastically larger than a random variable Y if $\Pr(X \le \beta) \le \Pr(Y \le \beta)$ for all β.[4] X is stochastically larger than Y if and only if $\mathbb{E}[\phi(X)] \ge \mathbb{E}[\phi(Y)]$ for all weakly increasing ϕ such that the expectation is defined. If X is stochastically larger than Y and $\mathbb{E}[\phi(X)] = \mathbb{E}[\phi(Y)]$ for some strictly increasing ϕ, then X and Y have the same distribution (Shaked and Shanthikumar 1994). The following lemma is then an immediate consequence of W_m being strictly increasing.

Lemma 1. *Suppose $\alpha_t(\mathbf{\Pi}_{-m})$ is stochastically larger than $\alpha_s(\mathbf{\Pi}_{-m})$. Then, agent m weakly prefers arriving in bin s. If agent m is indifferent between bins s and t, $\alpha_t(\mathbf{\Pi}_{-m})$ and $\alpha_s(\mathbf{\Pi}_{-m})$ have the same distribution.*

For $\mathbf{\Pi}^*$ to be a mixed-strategy Nash equilibrium, we must have for $m = 1, \ldots, M$,

$$V_m - \mathscr{W}_m(\mathbf{\Pi}^*) \ge V_m - \mathbb{E}[W_m(\alpha_t(\mathbf{\Pi}^*_{-m}) + 1)] \quad \text{for } t = 1, \ldots, T \qquad (6)$$

(see Fudenberg and Tirole 1996). This definition requires that an agent must weakly prefer randomizing over a set of actions to going to any one bin with certainty. Thus, she must be indifferent among any actions on which she puts positive probability.[5]

We now present an example that illustrates the nature of a mixed-strategy Nash equilibrium. It demonstrates that unlike a pure-strategy Nash equilibrium, a mixed-strategy Nash equilibrium may depend on the specific functions that characterize the agents' delay costs.

Example 1. Suppose $T > 2$, and $\lambda = \underline{\lambda} \ge 2$. (The latter implies that λ is an integer so $\tau = 0$.) Four agents are selected to randomize, placing equal weight on bins 1 and 2. For each bin $t > 2$, λ agents are selected and deterministically report to bin t. If agents remain, $\lambda - 2$ agents deterministically report to bin 1 and $\lambda - 2$ report to bin 2. Note that $\mathbb{E}[\alpha_t(\mathbf{\Pi})] = \lambda$ for all t. We claim this forms a mixed-strategy Nash equilibrium if all agents have expected congestion costs as given in (4) with $\theta_m = \theta$, i.e., $W(\alpha) = \theta - \theta/\alpha$.

[4] Note that we define the ordering in a weak sense.

[5] To see this classical result, multiply both sides of (6) by π^t_m and sum over t.

To establish this we need to verify that three types of agents are willing to follow the proposed equilibrium. Type 1 agents report deterministically to bins 1 or 2. Type 2 agents randomize between 1 and 2. Type 3 agents report deterministically to some bin $t > 2$. Let m_i denote a representative agent of each type for $i = 1, 2, 3$. First, consider a type 1 agent reporting deterministically to bin $j = 1$ or 2. Such an agent has no interest in moving to bin $3 - j$ by Lemma 1. Next, because $\alpha_t(\mathbf{\Pi}_{-m_1})$ is stochastically smaller than $\alpha_t(\mathbf{\Pi}_{-m_2})$ for all t, a type 2 agent has weakly higher delay costs from following the proposed equilibrium than a type 1 agent. Hence, if type 2 agents follow the equilibrium, so will type 1 agents.

For type 2 agents, $\mathbb{E}[\alpha_t(\mathbf{\Pi}_{-m_2})] = \lambda - 1/2$ for $t = 1, 2$. The agent prefers the proposed equilibrium to having $\alpha_t(\mathbf{\Pi}_{-m_2}) = \lambda - 1/2$ deterministically (because W is concave). Thus, she prefers participating in the equilibrium to deviating to $t > 2$ for which $\alpha_t(\mathbf{\Pi}_{-m_2}) = \lambda$.

Finally, a type 3 agent's cost from following the strategy is $W(\lambda)$ with certainty. Additionally, $\alpha_t(\mathbf{\Pi}_{-m_3}) - (\lambda - 2)$ has a binomial distribution with parameters $(4, 1/2)$. Therefore, if she deviates to bin 1 or 2, her costs increase by

$$\frac{W(\lambda-1) + 4W(\lambda) + 6W(\lambda+1) + 4W(\lambda+2) + W(\lambda+3)}{16} - W(\lambda)$$
$$= \theta(-9 - 2\lambda + 6\lambda^2 + 2\lambda^3)/2(\lambda-1)\lambda(\lambda+1)(\lambda+2)(\lambda+3),$$

which is positive because $\lambda \ge 2$. Hence, type 3 agents also follow the equilibrium.

While all pure-strategy Nash equilibria are independent of the assumed delay cost functions, this mixed-strategy Nash equilibrium depends on them in a crucial way. Agents are assumed to have a concave cost function and so behave in a risk-seeking fashion. Suppose, on the other hand, that a type 2 agent has a cost function as given in (3) with $\theta = 10$, i.e., $W_{m_2}(\alpha) = (10^\alpha - 1)/(9\alpha) - 1$. It is possible to show that such an agent always prefers to deviate to bin $t \ge 3$ for any $\lambda > 2$. Thus, the feasibility of an equilibrium depends critically on the delay cost function. In this case, the equilibrium can collapse if agents have convex costs and so act in a risk-averse manner. Given this observation, it is useful to consider whether any mixed-strategy Nash equilibrium can be independent of the agents' delay costs.

THEOREM 4. *A Nash equilibrium* $\mathbf{\Pi}^*$ *is independent of all agents' particular delay cost functions if and only if for* $m = 1, \ldots, M$:

(1) $\alpha_s(\mathbf{\Pi}^*_{-m})$ *is stochastically larger than* $\alpha_t(\mathbf{\Pi}^*_{-m})$ *for all* $s, t \in \{1, \ldots, T\}$ *such that* $\pi^t_m > 0$ *and* $\pi^s_m = 0$.

(2) *For all* $t, u \in \{1, \ldots, T\}$ *such that* $\pi^t_m > 0$ *and* $\pi^u_m > 0$,

$$\Pr(\alpha_t(\mathbf{\Pi}^*_{-m}) = \beta) = \Pr(\alpha_u(\mathbf{\Pi}^*_{-m}) = \beta) \tag{7}$$

for $\beta = 0, \ldots, M-1$.

PROOF. See the Appendix. □

The first condition of the theorem asserts that for a mixed-strategy Nash equilibrium to be independent of delay costs, an agent must randomize over what she perceives as less-crowded bins. However, the bins cannot be too different. One can show that if $\mathbf{\Pi}^*$ is a mixed-strategy Nash equilibrium for any W, we must have $\mathbb{E}[\alpha_t(\mathbf{\Pi}^*)] - \mathbb{E}[\alpha_s(\mathbf{\Pi}^*)] \leq 1$ for all $s, t = 1, \ldots, T$. Thus, a condition similar to that of Theorem 1 must hold in expectation. The second condition asserts that agents must view bins over which they randomize as identical. The following alters the previous example to be independent of delay cost functions.

EXAMPLE 2. Modify Example 1 so that only two agents randomize between bins 1 and 2 while $\lambda - 1$ agents report to those bins deterministically. Define types 1, 2, and 3 as before. From the perspective of a type 1 or 2, the distributions of $\alpha_1(\mathbf{\Pi}^*_{-m})$ and $\alpha_2(\mathbf{\Pi}^*_{-m})$ are identical, and the maximum value $\alpha_t(\mathbf{\Pi}^*_{-m})$ takes for $t = 1, 2$ is λ while $\alpha_s(\mathbf{\Pi}^*_{-m}) = \lambda$ for all $s > 2$. For a type 3 agent reporting to bin $t > 2$, $\alpha_t(\mathbf{\Pi}^*_{-m}) = \lambda - 1$, which is the smallest possible value of $\alpha_s(\mathbf{\Pi}^*_{-m})$ for $s = 1, 2$. Hence, this is an equilibrium for any W.

The above example shows that even if a mixed-strategy Nash equilibrium is independent of delay cost functions, it is not necessarily easier to implement than a pure-strategy Nash equilibrium. In this case, we must divide the agents into three groups, and within groups 1 and 3 we must assign individual agents to particular bins. It would be much simpler to have a symmetric Nash equilibrium that is independent of the agents' congestion cost function. In a symmetric Nash equilibrium, $\boldsymbol{\pi}^*_m = \boldsymbol{\pi}^*$ for all m, so $\mathbf{\Pi}^* = (\boldsymbol{\pi}^*, \ldots, \boldsymbol{\pi}^*)$.

LEMMA 2. *In a symmetric Nash equilibrium* $\mathbf{\Pi}^*$, *all elements of* $\boldsymbol{\pi}^*$ *must be strictly positive*.

PROOF. If $\pi^{t^*} = 0$ for some t, any agent could deviate to t and have no delay. □

We now define $\boldsymbol{\pi}^U = (1/T, \ldots, 1/T)$ and

$$\mathbf{\Pi}^U = (\boldsymbol{\pi}^U, \ldots, \boldsymbol{\pi}^U).$$

It is straightforward to show that $\mathbf{\Pi}^U$ is a mixed-strategy Nash equilibrium.

THEOREM 5. $\mathbf{\Pi}^U$ *is the only symmetric Nash equilibrium. It is independent of the agents' delay cost functions*.

PROOF. Suppose there is a symmetric Nash equilibrium $\mathbf{\Pi}^*$ distinct from $\mathbf{\Pi}^U$. It must have $\pi^{t^*} > \pi^{s^*}$ for some bins t and s. Note that $\alpha_t(\mathbf{\Pi}^*_{-m})[\alpha_s(\mathbf{\Pi}^*_{-m})]$ has a binomial distribution with parameters $M-1$ and π^{t^*} $[\pi^{s^*}]$. $\alpha_t(\mathbf{\Pi}^*_{-m})$ is then stochastically larger than $\alpha_s(\mathbf{\Pi}^*_{-m})$ for all m. However, because $\mathbf{\Pi}^*$ is a mixed-strategy Nash equilibrium and $\pi^{s^*} > 0$ (by Lemma 2), all agents must be indifferent between s and t. Hence, $\alpha_t(\mathbf{\Pi}^*_{-m})$ and $\alpha_s(\mathbf{\Pi}^*_{-m})$ must have the same distribution by Lemma 1, contradicting $\pi^{t^*} > \pi^{s^*}$. It is easy to verify that $\mathbf{\Pi}^U$ satisfies the requirements of Theorem 4. □

The equilibrium $\mathbf{\Pi}^U$ is an attractive alternative to other Nash equilibria because it requires minimal coordination. It is symmetric, independent of delay cost functions, and even independent of the number of agents. Thus, instead of full information, one needs to only assume that each agent knows just the number of bins (i.e., her possible actions).

In some sense, $\mathbf{\Pi}^U$ is an obvious outcome. If an agent is completely uninformed—unsure of how many others will seek service or of their delay costs—how could she do better than uniformly picking among the bins? It is important to recognize that such reasoning depends on strategically anticipating how others act. Uniformly randomizing only makes sense if others do so as well. If one conjectures that other agents are "early birds" or are prone to procrastination, uniformly picking an arrival time is no longer optimal. Thus, the arrival pattern generated by $\mathbf{\Pi}^U$ depends on the strategic interaction between the agents.

It is worth considering how societal costs under $\mathbf{\Pi}^U$ differ from costs under an arbitrary pure-strategy Nash equilibrium. If all agents have linear delay cost functions, they would be indifferent between implementing $\mathbf{\Pi}^U$ and some pure-strategy Nash equilibrium. There would be no reason to incur the cost

of running an appointment system (assuming the operating costs of the service facility are independent of variation in the arrival rate). On the other hand, if all agents have convex delay cost functions, randomization makes them worse off. Now an appointment system may be worthwhile if it is sufficiently cheap. This is especially true if facility operating costs are increasing in the variation of arrivals.

We now turn to our primary interest: How do arrivals under $\mathbf{\Pi}^U$ relate to arrival processes commonly assumed in the literature? We need the following definition.

Definition. The discrete-time arrival process $\{A(t)\colon t \in \{1, 2, \ldots, T\}\}$, where $A(t) = \sum_{i=1}^{t} \alpha_i$, is a *discrete-time Poisson process* with rate λ if and only if its per-period arrivals α_i are independent and identically Poisson distributed random variables with mean λ.

Clearly, as the sum of independent and identically distributed Poisson random variables, the cumulative number of arrivals $A(t)$ for any t is a Poisson random variable with parameter λt: for any integer k, $\Pr(A(t) = k) = (\lambda t)^k e^{-\lambda t}/k!$. An obvious way of creating a discrete-time Poisson process is to observe a standard continuous-time Poisson process at fixed intervals and map arrivals between observations to the start of the arrival interval. What is less obvious is how to create a continuous-time Poisson process from a discrete version.

Lemma 3. *Let $\{A(t)\colon t \in \{1, 2, \ldots, T\}\}$ be a discrete-time Poisson process with rate λ. Construct a continuous-time process $\{\hat{A}(t)\colon t \in [1, T+1]\}$ by assigning the kth arrival to bin t an arrival time of $t + u_{tk}$, where the random variables u_{tk} are independent and uniformly distributed on $[0, 1)$. $\{\hat{A}(t)\colon t \in [1, T+1]\}$ is a continuous-time Poisson process with rate λ.*

Proof. See the Appendix. □

Under $\mathbf{\Pi}^U$, arrivals to bin t have a binomial distribution with mean λ, and the resulting arrival pattern has some similarity to a Poisson process. Its increments are stationary. If one is told that over a subset of bins there have been k arrivals, then those arrivals are uniformly distributed across the bins. Where the similarity fails is independence. The joint distribution of arrivals is multinomial, and arrivals to distinct bins are not independent. However, we can relax this as the number of agents and time periods gets large.

Theorem 6. *Consider a sequence of systems (M_n, T_n) indexed by n such that for all n, M_n and T_n are integers, $M_n/T_n = \lambda$, $M_{n+1} > M_n$, and $M_n \to \infty$. Let $\mathbf{\Pi}_n^U$ denote the Nash equilibrium in system n in which all M_n agents play $\{1/T_n, \ldots, 1/T_n\}$.*

(1) The associated cumulative arrival process $\{A_n(t)\colon t \in \{1, 2, \ldots, T_n\}\}$ in system n converges in distribution to a discrete-time Poisson process as $n \to \infty$.

(2) Suppose that for all $n \geq \hat{n}$ one observes that $A_n(\hat{t}) = \hat{A} < M_{\hat{n}} - 1$ for some $\hat{t} < T_{\hat{n}}$. Let ν_n be the number of periods until the next arrival, i.e., $\alpha^n_{\hat{t}+j}(\mathbf{\Pi}_n^U) = 0$ for $j = 1, \ldots, \nu_n - 1$ and $\alpha^n_{\hat{t}+\nu_n}(\mathbf{\Pi}_n^U) > 0$. Then, $\lim_{n\to\infty} P(\nu_n \leq k \mid A_n(\hat{t}) = \hat{A}) = 1 - e^{-k\lambda}$ for $k = 1, 2, \ldots$.

Proof. See the Appendix. □

Thus, agents under Nash equilibrium $\mathbf{\Pi}^U$ generate an arrival pattern that approaches a discrete-time Poisson process, i.e., arrivals in disjoint intervals are independent and identically distributed Poisson random variables. Further, the time between bins with positive arrivals goes to a discrete version of the exponential distribution. Hence, insights on managing service facilities generated from models assuming renewal interarrivals remain valid even if customers strategically choose when to seek service. If the number of customers and arrival bins is large, the arrival pattern that results from the most plausible Nash equilibrium is well approximated by a discrete-time Poisson process.[6]

We emphasize two points. First, Theorem 6 does not depend on the ample capacity assumption. As long as $\mathbf{\Pi}_n^U$ is a Nash equilibrium for all systems in the sequence, arrivals will converge to a discrete-time Poisson process. We exploit this below. Second, the theorem does depend on the strategic interactions between customers. A given agent chooses an arrival point in such a way that total arrivals form a Poisson process because of how the other agents play. Not all mixed-strategy Nash equilibria lead to Poisson arrivals. Consider Example 2. If one scales up that example, the equilibrium continues to hold but does

[6] Theorem 6 can be seen as extending the well-known result on the convergence of a binomial to a Poisson random variable to a multinomial random variable and a Poisson process. Feller (1957) presents a problem on approximating a multinomial distribution by independent Poisson random variables, but he does not relate this to an arrival process.

not lead to Poisson arrivals. Our result also goes beyond the usual interpretation that a Poisson process arises from having a large number of agents acting independently. Yes, $\mathbf{\Pi}^U$ assumes that each agent picks an arrival bin independently of all others, but the same is true for any mixed-strategy Nash equilibrium and not all mixed-strategy Nash equilibria lead to Poisson arrivals. Not only does Nash equilibrium $\mathbf{\Pi}^U$ lead to Poisson arrivals, but it also has a number of other appealing properties. It is the only symmetric equilibrium and is independent of the delay cost functions and the number of agents. Hence, it requires very little coordination to implement.

Theorem 6 is a limiting result, so we now consider how quickly $\Pr(\alpha_t(\mathbf{\Pi}_n^U) = k)$ converges to a Poisson distribution for two arrival rates ($\lambda = 2$ and $\lambda = 0.6$). We begin with $T_1 = 25$ and then increase the number of bins and agents holding λ constant. Table 2 reports the maximum absolute deviation in the probability mass function (PMF) and the cumulative distribution function (CDF) for each iteration. The lower arrival rate converges more slowly but is well approximated by a Poisson distribution; with 50 bins and 30 agents, the maximum deviation is less than half a percent.

Table 3 examines how quickly the dependence between bins diminishes. We assume $M = 400$ and $T = 200$ (so $\lambda = 2$) and examine the distribution of $\alpha_{t+1}(\mathbf{\Pi}^U)$ conditional on observing bins 1 through t for $t = 50$, 100, and 150. As the conditional distribution depends only on the cumulative number of arrivals, we consider having arrivals run 10 and 20 agents above or below their expected value. We report the maximum absolute deviation between the PMF and the CDF of $\alpha_{t+1}(\mathbf{\Pi}^U)$ given $A(t)$ and a Poisson distribution with $\lambda = 2$. Thus, if one observes 100 bins and sees total arrivals of 210, we compare a Poisson distribution with $\lambda = 2$ to a binomial with parameters $(400 - 210)$ and $1/(200 - 100)$. As one would expect, the fit is not as close as for the unconditional distribution, but it remains a reasonable approximation as long as one does not know "too much." When one has observed 150 bins (75% of the horizon) and the arrivals deviate significantly from the average, the Poisson is not a close fit (the maximum deviation is over 5%), but when observing fewer bins and having smaller deviations from the mean, the fit is much tighter.

Table 2 Unconditional Covergence to a Poisson Distribution

	Maximum Deviation	
(M, T)	PMF	CDF
(50, 25)	0.00556	0.00552
(100, 50)	0.00274	0.00273
(200, 100)	0.00136	0.00136
(400, 200)	0.00068	0.00068
(15, 25)	0.00952	0.00673
(30, 50)	0.00468	0.00333
(60, 100)	0.00232	0.00165
(120, 200)	0.00116	0.00083

Table 3 Conditional Convergence to a Poisson Distribution

Observed Number of Bins	Deviation from Mean Arrivals	Maximum Deviation	
		PMF	CDF
50	−20	0.01811	0.03616
50	−10	0.00960	0.01868
50	10	0.00908	0.01810
50	20	0.01834	0.03645
100	−20	0.02715	0.05423
100	−10	0.01417	0.02781
100	10	0.01365	0.02718
100	20	0.02847	0.05563
150	−20	0.05384	0.10802
150	−10	0.02741	0.05450
150	10	0.02745	0.05450
150	20	0.06331	0.11697

To formalize the comparison between $\alpha_t(\mathbf{\Pi}_n^U)$ and a Poisson random variable, we simulated 1,000 draws of $\alpha_t(\mathbf{\Pi}_n^U)$ and tested whether one could reject the null hypothesis that the resulting output was in fact generated by a Poisson random variable with the appropriate mean using a chi-square goodness of fit test (Larsen and Marx 1986). For both examples in Table 2, we could not reject this null hypothesis at a 95% confidence level even for $T_1 = 25$. Similar tests for the examples in Table 3 show that one cannot reject the Poisson distribution (with $\lambda = 2$) as generating arrivals even when one has observed 20 arrivals above or below the mean over 50 bins and 10 arrivals above or below the mean over 100 bins. However, when one has observed 150 bins or when one has observed 20 arrivals above or below the mean over 100 bins, one can reject the hypothesis that arrivals are in fact Poisson.

3.3. Extensions

We briefly consider some extensions for which $\mathbf{\Pi}^U$ continues to be an equilibrium.

3.3.1. A Finite Waiting Room. We have thus far assumed all agents arriving to bin t can enter and be served. We now suppose that the system can only hold K customers. Customers are randomly sequenced upon arrival. The first K are admitted and served according to the established sequence. Remaining customers are denied service. If agent m is admitted, she incurs a cost $C_m(k)$ if she is the kth person processed. Assume $V_m \geq C_m(K)$ so she values service if admitted. If she is denied admission, she incurs a cost $\omega_m \geq 0$ and does not receive V_m. Her expected net benefit given α_t arrivals is then

$$U_m(\alpha_t) = \begin{cases} (1/\alpha_t)\sum_{k=1}^{K}[V_m - C_m(k)] & \text{if } \alpha_t \leq K, \\ (K/\alpha_t)\big((1/K)\sum_{k=1}^{K}[V_m - C_m(k)]\big) & \\ \qquad -((\alpha_t - K)/\alpha_t)\omega_m & \text{if } \alpha_t > K. \end{cases}$$

This net benefit function models a system with a finite waiting room and first in first out (FIFO) service.[7] An agent denied entry receives no net benefit from service and may incur a loss. One could specify a similar cost function for the batch model of (4). As $U_m(\alpha_t)$ decreases in α_t, our analysis is unchanged, and $\mathbf{\Pi}^U$ is still an equilibrium.

3.3.2. Exogenous Arrivals. Suppose that in each bin there is a stochastic shock in addition to the arrival of the agents. Let $X = (X_1, \ldots, X_T)$ denote the vector of shocks. Such a shock could be additional customers arriving from some other source. Agents choose their bins before the shocks are observed. If the realized number of agents is α_t, then agent m's congestion costs are $W_m(\alpha_t + x_t)$, where x_t is the realized value of X_t. If the marginal distribution of the shocks is the same for all bins, our analysis goes through. When considering pure strategies, agents spread themselves out as much as possible. For mixed strategies, $\mathbf{\Pi}^U$ is again viable. Note that we require the marginal distributions to be the same but do not require independence. Thus, we could have the arrivals in bins being positively or negatively correlated.

[7] Alternatively, assume that $V_m \geq C_m(K)$ but $V_m < C_m(K+1)$ for all m and allow agents to balk after observing where they will be sequenced among the α_t arrivals. The system will then function as if it had a finite waiting room. See Naor (1969).

3.3.3. Priorities. Suppose there are two classes of agents. There are M_1 type 1 agents and M_2 type 2 agents. Let α_t^i be the number of type i agents arriving to bin t. Suppose m_1 is a type 1 agent with congestion costs $W_{m_1}(\alpha_t^1, \alpha_t^2)$. We assume that

$$\frac{\partial W_{m_1}(\alpha_t^1, \alpha_t^2)}{\partial \alpha_t^1} > \frac{\partial W_{m_1}(\alpha_t^1, \alpha_t^2)}{\partial \alpha_t^2} = 0.$$

Let m_2 be a type 2 agent with congestion costs $W_{m_2}(\alpha_t^1, \alpha_t^2)$. We assume that

$$\frac{\partial W_{m_2}(\alpha_t^1, \alpha_t^2)}{\partial \alpha_t^1} \geq \frac{\partial W_{m_2}(\alpha_t^1, \alpha_t^2)}{\partial \alpha_t^2} > 0.$$

These cost functions are consistent with a priority scheme that serves type 1 arrivals before type 2 arrivals. A type 1's net utility is unaffected by the arrival of a type 2 customer, but type 1 arrivals always lower the utility of a type 2 customer.

First, consider type 1 agents. Because their waits are independent of the actions of the second type, they face the problem analyzed above. Suppose all type 1 agents play $\mathbf{\Pi}^U$, and consider type 2 customers. If type 2 agents also play $\mathbf{\Pi}^U$, agent m_2's expected costs from reporting to any bin t are $\mathbb{E}[W_m(\alpha_t^1(\mathbf{\Pi}^U_{-m_2}), \alpha_t^2(\mathbf{\Pi}^U_{-m_2}) + 1)]$. Because her expected costs are the same for all t, agent m_2 is also willing to follow $\mathbf{\Pi}^U$.

3.3.4. Multiple Servers. The waiting cost formulation (1) implicitly assumes a single server. Suppose instead that there are $N \geq 1$ servers. For all m, $C_m(k) = 0$ for all $k \leq N$ but $C_m(k) > 0$ for $k > N$. W_m is no longer strictly increasing so one can have a pure-strategy Nash equilibrium $\mathbf{\Pi}^*$ with $\alpha_t(\mathbf{\Pi}^*) - \alpha_s(\mathbf{\Pi}^*) > 1$ (if $N > \underline{\lambda} + 1$). However, $\mathbf{\Pi}^U$ remains a Nash equilibrium (although it is not the only symmetric equilibrium if $N \geq M$). Each agent perceives all $\alpha_t(\mathbf{\Pi}^U_{-m})$ as being identically distributed and, hence, is willing to randomize among them.

4. Time-Varying Preferences

We now allow the value of service to depend on the period in which service is received. Divide the horizon into two sets, H and L. Agents are symmetric and value receiving service in a bin falling in H at V_H. Receiving service in L is valued at V_L with $V_H > V_L$. We impose no restrictions on the location of the "sweet spot" of the horizon. The high-value bins

may be at the start of the horizon (e.g., getting to the gym before work), at the end of horizon (e.g., running weekend errands after sleeping in), or in the middle of the horizon (e.g., going to lunch neither too early nor too late). Let $T_H = \|H\|$ denote the number of bins in set H and $T_L = \|L\| = T - T_H$ denote the number of bins in set L. Let $\eta = T_H/T$. Agents have an identical linear cost function as given in (2), with $\theta_m = \theta$ for all m.[8] Hence,

$$W_m(\alpha) = W(\alpha) = \frac{1}{\alpha}\sum_{i=1}^{\alpha}\theta(k-1) = \theta\frac{\alpha-1}{2}.$$

We consider symmetric Nash equilibria. Suppose each agent puts probability on $1/T_H$ on selecting a bin in H while putting zero probability on any bin in L. Let $\mathbf{\Pi}^H$ denote the resulting strategy profile. Agent m's expected pay off from reporting to bin $h \in H$ is

$$\begin{aligned} V_H - \mathbb{E}[W(\alpha_h(\mathbf{\Pi}^H_{-m}) + 1)] &= V_H - \frac{\theta}{2}\frac{M-1}{T_H} \\ &= V_H - \frac{\theta}{2\eta}\left(\lambda - \frac{1}{T}\right). \end{aligned}$$

If she deviates to bin $l \in L$, she receives V_L with no delay cost. $\mathbf{\Pi}^H$ is an equilibrium if

$$\frac{2\eta}{\theta}(V_H - V_L) + \frac{1}{T} \geq \frac{M}{T} = \lambda. \tag{8}$$

Thus, Lemma 2 does not necessarily hold. However, if a symmetric Nash equilibrium puts positive weight on any bin in a set, it must put positive weight on all bins in that set. Given the assumed cost function, bins in the same set must all have the same expected number of arrivals. Consequently, in a symmetric Nash equilibrium all agents must put weight π_H on each bin in H and π_L on each bin in L. Additionally, $\pi_L > 0$ if (8) fails, and

$$T_H\pi_H + T_L\pi_L = 1. \tag{9}$$

Let $\mathbf{\Pi}^o$ denote the resulting strategy profile. For $\mathbf{\Pi}^o$ to be a Nash equilibrium, agent m must be indifferent between bins $h \in H$ and $l \in L$:

$$V_H - \mathbb{E}[W(\alpha_h(\mathbf{\Pi}^o_{-m}) + 1)] = V_L - \mathbb{E}[W(\alpha_l(\mathbf{\Pi}^o_{-m}) + 1)].$$

[8] Having θ vary over the horizon yields results similar to those presented below.

Because $\mathbb{E}[\alpha_h(\mathbf{\Pi}^o_{-m})] = \pi_H(M-1)$ and $\mathbb{E}[\alpha_l(\mathbf{\Pi}^o_{-m})] = \pi_L(M-1)$, we have

$$\pi_H - \pi_L = \frac{2(V_H - V_L)}{\theta(M-1)}.$$

Because $V_H > V_L$, a representative high value bin will have a higher equilibrium arrival rate than a representative low value bin, and the difference in arrival rates is sufficiently high to dissipate any gains from receiving service in bin H. Agents do not anticipate a higher net utility in set H because those bins are so congested. Using (9), we then have

$$\begin{aligned} \pi_H(M, T) &= \frac{1}{T} + \frac{2(1-\eta)}{\theta(M-1)}(V_H - V_L), \\ \pi_L(M, T) &= \frac{1}{T} - \frac{2\eta}{\theta(M-1)}(V_H - V_L). \end{aligned}$$

For both π_H and π_L, the second term represents the deviation from the uniform Nash equilibrium $\mathbf{\Pi}^U$. This deviation increases as the gain from being in the sweet spot of the horizon ($V_H - V_L$) increases. It falls if either crowding is likely (M is large) or waiting is very costly (θ is high). π_H and π_L both decrease as η increases. High-value bins are inherently less crowded as more bins are added to H; low-value bins must also become less crowded (despite there being relatively fewer of them) to continue to attract agents.

These comparative statics have some managerial implications. Suppose management responds to time-varying arrivals by taking actions to increase V_H (e.g., live piano music during busy hours) or lower congestion costs in the favorable part of the horizon (e.g., adding capacity). These results suggest that difference between arrival rates to the sets will increase, i.e., attempts to deal with time-varying arrivals may exacerbate the swings in arrivals.

The symmetric Nash equilibrium $\mathbf{\Pi}^o$ for the time-dependent rewards case differs from the symmetric Nash equilibrium $\mathbf{\Pi}^U$ for the stationary values case in significant ways. $\mathbf{\Pi}^U$ is independent of any agent's cost function and the total number of agents. $\mathbf{\Pi}^o$ depends on both. However, one can still examine the limiting case as the number of bins gets large.

THEOREM 7. *Consider a sequence of system* (M_n, T_n) *such that for all* n, M_n *and* T_n *are integers,* $M_n/T_n = \lambda$, $M_{n+1} > M_n$, *and* $\lim_{n\to\infty} M_n = \infty$. *For each* n, *agents*

receive reward V_H in set H_n and reward V_L in set L_n, where $\|H_n\| = \eta T_n$ and $\|L_n\| = (1-\eta)T_n$. Assume (8) *does not hold for $n = 1$. Let $\mathbf{\Pi}_n^o$ denote the equilibrium in which all M_n agents put probability $\pi_H^n = \pi_H(M_n, T_n)$ on each bin in H_n and probability $\pi_L^n = \pi_L(M_n, T_n)$ on each bin in L_n.*

(1) *For $h \in H_n$, $\lim_{n\to\infty} \Pr(\alpha_h(\mathbf{\Pi}_n^o) = k) = e^{-\lambda_H}\lambda_H^k/k!$, where $\lambda_H = \lambda + (2/\theta)(1-\eta)(V_H - V_L)$.*

(2) *For $l \in L_n$, $\lim_{n\to\infty} \Pr(\alpha_l(\mathbf{\Pi}_n^o) = k) = e^{-\lambda_L}\lambda_L^k/k!$, where $\lambda_L = \lambda - (2/\theta)\eta(V_H - V_L)$.*

(3) *$\lim_{n\to\infty} \Pr(\alpha_t(\mathbf{\Pi}_n^o) = j, \alpha_s(\mathbf{\Pi}_n^o) = k) = (e^{-\lambda_t}\lambda_t^j/j!)\cdot(e^{-\lambda_s}\lambda_s^k/k!)$, where λ_t equals λ_H if $t \in H_n$ and λ_L otherwise and λ_s is defined similarly.*

Proof. Substituting $T_n = M_n/\lambda$, one has

$$\pi_H^n = \frac{1}{M_n}\left(\lambda + \frac{2}{\theta(1-1/M_n)}(1-\eta)(V_H - V_L)\right),$$

$$\pi_L^n = \frac{1}{M_n}\left(\lambda - \frac{2}{\theta(1-1/M_n)}\eta(V_H - V_L)\right).$$

The rest of the proof is then similar to Theorem 6. □

Competing customers with nonstationary preferences generate a nonstationary discrete-time Poisson process. Arrivals are higher in the high-value bins, and competition eliminates any benefit from selecting a high-value bin; the sweet spot of the horizon is sufficiently crowded that all agents are indifferent between reporting to any bin in H and any bin in L.

To extend this approach to $R > 2$ levels of rewards, assume that all agents put positive probability on R sets of bins. The agents must then be indifferent between the highest-value set and the next $R-1$ highest-value sets. Together with the appropriate analog of (9), this leads to a set of R linear equations in R unknowns. Within each set, arrivals to a given bin will have a binomial distribution, and higher-value bins will have higher arrival rates.

5. Equilibria with Limited Capacity

We now return to the case of time-invariant preferences but suppose the system can only serve a limited number of customers in each period. Customers unserved in period t carry over to period $t+1$. Our intention is to develop conditions such that $\mathbf{\Pi}^U$ is again a Nash equilibrium so Theorem 6 continues to hold and a discrete-time Poisson process is still a valid approximation for the arrival process generated by strategic customers.

Let I_t denote the inventory of customers in the system at the start of period t prior to any new arrivals. Let s_t denote the maximum number of customers that can be served in period t. In each period, I_t agents are carried over from period $t-1$, α_t new customers arrive, s_t is realized, and $\min\{I_t + \alpha_t, s_t\}$ customers exit the system. The number of customers carried into period $t+1$ is therefore

$$I_{t+1} = [I_t + \alpha_t - s_t]^+,$$

where $[x]^+ = \max\{x, 0\}$. We assume s_t is a nonnegative random variable that takes only integer values. The draw in each period is identically and independently distributed (IID). $\mathbb{E}[s_t] = \mu > \lambda$ and $\Pr(s_t < M) > 0$. The latter implies limited capacity; there is some chance that the system cannot process all arrivals. Restricting s_t to be integer valued simplifies the state space. I_t is always integer valued, and we only need to track the number of agents in the system as opposed to the work in the system. Possible examples of s_t include a Poisson random variable with mean μ or a Bernoulli random variable with probability of success equal to μ. Note that if we take I_t as the state of the system, it is not a Markov chain because the distribution of α_t depends on the entire history of the process.[9]

We assume a FIFO discipline. Arrivals in period t must wait for the I_t customers already in the system. If $\alpha_t \geq 2$, these new arrivals are randomly ordered and served in that sequence. Delay costs thus follow a modified form of (1):

$$W_m(\alpha_t, I_t) = \frac{1}{\alpha_t}\sum_{k=1}^{\alpha_t} C_m(k + I_t).$$

Let I_1 denote the initial population of customers. I_1 is a random variable such that $\Pr(I_1 = k) = f_1(k)$ for $k = 0, 1, \ldots$. The agents know the distribution of I_1 but do not see its realized value until after they have selected their arrival bin. Let I_{T+1} denote the number of customers who remain unserved at the end of the horizon. We assume the system continues to operate until all customers have been served. Additional

[9] The distribution of α_t given the process history depends on how many bins $(T+1-t)$ and agents $(M - A(t-1))$ remain. Hence, $(t, A(t-1), I_t)$ is a Markov chain.

service draws $s_{T+1}, s_{T+2}, \ldots$ are taken until all I_{T+1} customers have exited the system where draw s_{T+k} has the same distribution as s_t for $t = 1, \ldots, T$. For example, if $s_t = 1$ with certainty, an additional I_{T+1} periods of operation are required. Assuming continued operations assures the system will eventually clear and removes the end of horizon effect. Together with the FIFO service discipline, it assures that $W_m(\alpha_T, I_T) = W_m(\alpha_t, I_t)$ for $t < T$ as long as $(\alpha_T, I_T) = (\alpha_t, I_t)$.

Agent m's objective is again to maximize her net benefit $U_m(\alpha_t, I_t)$ or equivalently to minimize her expected delay cost $W_m(\alpha_t, I_t)$. Agents still choose their arrival bins simultaneously. The novel aspect of imposing limited capacity is that an agent's cost now depends on both the number of agents that arrive with her as well as the existing inventory of agents. Because of the FIFO discipline, agent m's costs depend only on the number of agents that arrive before her and with her, not on the number that come after her.

Let $\mathbf{\Pi}^*$ be a candidate Nash equilibrium. $\alpha_t(\mathbf{\Pi}^*_{-m})$ and $I_t(\mathbf{\Pi}^*_{-m})$ respectively denote the number of arrivals to bin t and the number of agents already present in bin t under $\mathbf{\Pi}^*$ holding agent m out. Because $I_1(\mathbf{\Pi}^*_{-m}) = I_1$, $\mathbf{\Pi}^*$ also depends on the distribution of I_1. We suppress this dependence to simplify the notation. For $\mathbf{\Pi}^*$ to be a Nash equilibrium, we require for $m = 1, \ldots, M$,

$$V_m - \mathcal{W}_m(\mathbf{\Pi}^*) \geq V_m - \mathbb{E}[W_m(\alpha_t(\mathbf{\Pi}^*_{-m}) + 1, I_t(\mathbf{\Pi}^*_{-m}))] \quad \text{for } t = 1, \ldots, T, \quad (10)$$

where

$$\mathcal{W}_m(\mathbf{\Pi}^*) = \sum_{t=1}^{T} \boldsymbol{\pi}_m^t \mathbb{E}[W_m(\alpha_t(\mathbf{\Pi}^*_{-m}) + 1, I_t(\mathbf{\Pi}^*_{-m}))].$$

Lemma 4. *Given M, T, and the distributions of s_t and I_1, a Nash equilibrium $\mathbf{\Pi}^*$ exists.*

Proof. Every finite strategic form game has a mixed-strategy Nash equilibrium. See Theorem 1.1 of Fudenberg and Tirole (1996). □

The lemma does not guarantee that $\mathbf{\Pi}^U$ is a Nash equilibrium. Indeed, it may not be. Suppose $I_1 \equiv 0$ and that all agents follow $\mathbf{\Pi}^U$. If agent m reports to bin 1, she expects arrivals of $\alpha_t(\mathbf{\Pi}^U_{-m})$ but no inventory of existing customers. In bin 2, she again expects arrivals of $\alpha_t(\mathbf{\Pi}^U_{-m})$ but now $\Pr(I_2 > 0) > 0$ (because $\Pr(s_t < M) > 0$). She consequently strictly prefers the first bin, and $\mathbf{\Pi}^U$ cannot be a Nash equilibrium. See Glazer and Hassin (1983).

For $\mathbf{\Pi}^U$ to be an equilibrium, we need an alternative initialization. Consider the following:

$$\hat{I}_{t+1} = [\hat{I}_t + \hat{\alpha}_t - s_t]^+,$$

where s_t is as defined above and $\hat{\alpha}_t$ has a binomial distribution with parameters $M - 1$ and $1/T$. Draws of s_t and $\hat{\alpha}_t$ are independent across periods. While the inventory process for our system I_t does not form a Markov chain, $\hat{I}_t$ does. Let $\hat{P}_{ij} = \Pr(\hat{I}_{t+1} = j \mid \hat{I}_t = i)$. We assume that $\hat{I}_t$ is ergodic and denote its stationary distribution by $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \ldots)$.[10] That is, $\gamma_j = \sum_{i=0}^{\infty} \gamma_i \hat{P}_{ij}$. (See Ross 1983.) It turns out that $\boldsymbol{\gamma}$ is the initialization we need.

Theorem 8. *If the distribution of I_1 is such that $f_1(k) = \gamma_k$ for $k = 0, 1, \ldots$, then $\mathbf{\Pi}^U$ is a Nash equilibrium for any set of agent delay cost functions.*

Proof. Given that others play $\mathbf{\Pi}^U$, agent m perceives the distribution of $\alpha_t(\mathbf{\Pi}^U_{-m})$ as independent of t. If m deviates from $\mathbf{\Pi}^U$, it must be because of the evolution of the inventory process $I_t(\mathbf{\Pi}^U_{-m})$. Suppose the realized value of I_1 equals i. We have

$$I_2(\mathbf{\Pi}^U_{-m}) = [i + \alpha_1(\mathbf{\Pi}^U_{-m}) - s_t]^+.$$

Because $\alpha_1(\mathbf{\Pi}^U_{-m})$ has a binomial distribution with parameters $M - 1$ and $1/T$, we have that $\Pr(I_2(\mathbf{\Pi}^U_{-m}) = j \mid I_1 = i) = \hat{P}_{ij}$ and $\Pr(I_2(\mathbf{\Pi}^U_{-m}) = j) = \sum_{i=0}^{\infty} \gamma_i \hat{P}_{ij} = \gamma_j$. I_1 and I_2 then have the same distribution. An induction extends the result to any t. $\mathbf{\Pi}^U$ is thus a Nash equilibrium because $\mathbb{E}[W_m(\alpha_t(\mathbf{\Pi}^U_{-m}) + 1, I_t(\mathbf{\Pi}^U_{-m}))]$ is independent of t for any W_m. □

It is tempting to interpret Theorem 8 as saying that if the system starts in steady state, the agents play $\mathbf{\Pi}^U$, but this is not quite correct. First, I_t is not a Markov chain. Second, even if one considers a Markov version of I_t in which arrivals are independent across periods, its transition probabilities would not be $\hat{P}_{ij}$ because its arrivals would have a different distribution than $\hat{\alpha}_t$. That said, as M and T get large,

[10] Note that $\hat{I}_{t+1}$ is a discrete-time queue that is ergodic because we have assumed $\mu > \lambda$.

the distribution of $\hat{\alpha}_t$ converges to that of $\alpha_t(\mathbf{\Pi}^U)$, and $\boldsymbol{\gamma}$ converges to the steady-state distribution of I_t. Thus, looking at a limiting system as in Theorem 6, we have that Poisson arrivals see time averages of the state of the system.

A steady-state initialization may seem a limitation, but it is sufficient for our purposes. We are interested in whether previous work was limited by assuming renewal arrivals while ignoring that customers may pick arrival times strategically. Most existing academic research (e.g., Mendelson 1985) and managerial literature (e.g., Cleveland and Mayben 1997) also assume that arrivals experience steady-state waits. Thus, if one believes that it is sufficient to look at long-run average waits, our results suggest that strategic customers will plausibly produce an arrival pattern that approaches a discrete-time Poisson process.

Additionally, strategic customers will, in a sense, take the system to steady state. Specifically, they will equalize delay costs over the horizon. This is inherent in the definition of a Nash equilibrium given in (10). If agent m puts a positive probability on multiple bins, she must expect the same waiting costs in those bins. Suppose all agents have linear delay cost as in (2) and that for some initialization of I_t (not necessarily $\boldsymbol{\gamma}$) there exists a Nash equilibrium $\mathbf{\Pi}'$ such that some agent m puts positive probability on bins s and t. It must be

$$\mathbb{E}\big[\alpha_t(\mathbf{\Pi}'_{-m}) - \alpha_s(\mathbf{\Pi}'_{-m})\big] = 2\big(\mathbb{E}[I_s(\mathbf{\Pi}'_{-m}) - I_t(\mathbf{\Pi}'_{-m})]\big).$$

Differences in arrival rates compensate for differences in inventory levels, and equilibrium workload levels across bins must be constant. If a bin is expected to have a low inventory (as with $I_1 \equiv 0$), it must have a higher arrival rate and vice versa. In general, *any* Nash equilibrium will require that agents expect a (nearly) constant delay cost across the horizon, so equilibrium play will result in arrival rates that eliminate differences across bins.

This last point relies on time-invariant preferences. Combining time-varying preferences with limited capacity requires multiple initializations. For example, if the low-value bins occur at the start of the horizon, one would need to inject additional inventory in the system at the start of the high-value period.

6. Conclusion

We have presented a simple timing game. A set of customers seek service over some horizon. All find congestion costly and so try to arrive when the facility is underutilized. Working in discrete time, we characterize pure-strategy Nash equilibria for the case of ample capacity. Agents try to spread out as much as possible. Symmetric agents with well-behaved delay cost functions choose a socially efficient outcome that minimizes total waiting costs.

While potentially efficient, pure-strategy equilibria are difficult to implement in this setting. We consequently identify a unique symmetric Nash equilibrium. This equilibrium is independent of both the delay cost functions and the number of agents. Further, as the number of agents and time periods get large, the number of arrivals in any period goes to a Poisson distribution and the number of arrivals across bins becomes independent. Thus, a large population of strategic customers seeking to avoid congestion generates an arrival pattern well approximated by a discrete-time Poisson process. Our results extend to the case of limited capacity given an appropriate initialization of the system. If customer valuations of service vary over the horizon, one has a nonstationary discrete-time Poisson process.

Our model lends support to the traditional literature on the management of service systems. This work has generally assumed that customer arrival times are governed by a renewal process while ignoring the possibility that customers strategically try to avoid congestion. We show that assuming renewal interarrival times is acceptable given a large population and long horizon. In addition, our model offers some insights on when a reservation system may be worthwhile and how customers react to time-dependent service values.

While discrete time is a reasonable approximation of human behavior, traditional queuing models are in continuous time. Lemma 3 suggests a possible way to convert our discrete-time process to a continuous-time Poisson process. Suppose agents must choose from the discrete set $\{1, \ldots, T\}$ but their actual arrival times are subject to independent shocks; when agent m targets arriving at time t, her actual arrival time is uniformly distributed on $[t, t+1)$. (Ostrovsky and Schwarz 2003 employ a similar approach.) If others

play $\mathbf{\Pi}^U$, agent m perceives all bins as identical so she too will be willing to follow the equilibrium.

More challenging is to assume that agents choose arrival times from a continuous set. Consider the case of ample capacity and suppose each agent picks an arbitrary strictly continuous distribution on $[0, T]$. The chance of having simultaneous arrivals is zero, so any set of strictly continuous distributions is a Nash equilibrium. Thus, the ample capacity case is only interesting in discrete time. To work in continuous time, one must immediately worry about the service dynamics of the system as in Glazer and Hassin (1983).

We note that the discrete-time ample-capacity case is open to multiple interpretations. Rather than picking discrete arrival bins, agents could be determining which of many simultaneous markets or stores to visit. Faced with a large number of identical choices (e.g., an infinite number of Starbucks coffee shops), agents plausibly implement a Nash equilibrium that results in each store seeing arrivals closely resembling independent Poisson draws. We feel that this is a promising framework for examining interesting phenomena in services because it captures the most relevant aspect of services (e.g., degradation of the service experience due to congestion) while suppressing the complications of queuing dynamics.

Appendix

Proof of Corollary 1. We use an induction on T. Fix $T = 2$ and select a bin to have $\underline{\lambda}$ agents.

$$\binom{M}{\underline{\lambda}} = \frac{M!}{\underline{\lambda}!^{(T-\tau)}(\underline{\lambda}+1)!^{\tau}}$$

gives the number of equilibria given $\underline{\lambda}$ agents are assigned to the selected bin. $\binom{T}{\tau}$ then accounts for the number of ways to select τ bins to have $\underline{\lambda}+1$ agents. Suppose the result holds for $T-1$ bins and consider the case of T bins. Fix the τ bins assigned $\underline{\lambda}+1$ agents. There must be a bin with only $\underline{\lambda}$ agents. Assign $\underline{\lambda}$ agents to the earliest such bin. The number of equilibria vectors satisfying these criteria are

$$\binom{M}{\underline{\lambda}} \frac{(M-\underline{\lambda})!}{\underline{\lambda}!^{(T-1-\tau)}(\underline{\lambda}+1)!^{\tau}} = \frac{M!}{\underline{\lambda}!^{(T-\tau)}(\underline{\lambda}+1)!^{\tau}}.$$

The term $\binom{M}{\underline{\lambda}}$ accounts for the $\underline{\lambda}$ agents assigned to the earliest time period with only $\underline{\lambda}$ arrivals. The next term relies on the inductive hypothesis. The result follows. □

Proof of Theorem 4. Suppose both conditions hold. By Lemma 1, agent m weakly prefers t to s for all W_m. By (7), she is indifferent among bins over which she randomizes for all W_m. $\mathbf{\Pi}^*$ is a Nash equilibrium for all W_m. Now suppose $\mathbf{\Pi}^*$ is a Nash equilibrium for all W_m and that (7) fails for some agent m and some bins t and u. There must exist $\beta < \beta'$ such that

$$\Pr(\alpha_t(\mathbf{\Pi}^*_{-m}) = \beta) > \Pr(\alpha_u(\mathbf{\Pi}^*_{-m}) = \beta),$$

$$\Pr(\alpha_t(\mathbf{\Pi}^*_{-m}) = \beta') < \Pr(\alpha_u(\mathbf{\Pi}^*_{-m}) = \beta').$$

Because m randomizes over t and u, they must result in the same expected congestion cost

$$\mathbb{E}[W_m(\alpha_t(\mathbf{\Pi}^*_{-m}) + 1)] = \mathbb{E}[W_m(\alpha_u(\mathbf{\Pi}^*_{-m}) + 1)].$$

Pick a W_m such that this equality holds. Because W_m is strictly increasing, we can create a new cost function $\widetilde{W}_m$ such that $\widetilde{W}_m(\beta+1) = W_m(\beta+1) + \varepsilon$, $\widetilde{W}_m(\beta'+1) = W_m(\beta'+1) - \varepsilon$, and $\widetilde{W}_m = W_m$ otherwise. We must then have $\mathbb{E}[\widetilde{W}_m(\alpha_t(\mathbf{\Pi}^*_{-m}) + 1)] > \mathbb{E}[\widetilde{W}_m(\alpha_u(\mathbf{\Pi}^*_{-m}) + 1)]$, so the equilibrium is not independent of the cost function.

Now suppose condition 1 fails. There must exist an integer $\hat{\beta}$ and agent m such that $\Pr(\alpha_s(\mathbf{\Pi}^*_{-m}) \le \hat{\beta} - 1) > \Pr(\alpha_t(\mathbf{\Pi}^*_{-m}) \le \hat{\beta} - 1)$. Equivalently, $\Pr(\alpha_s(\mathbf{\Pi}^*_{-m}) \ge \hat{\beta}) < \Pr(\alpha_t(\mathbf{\Pi}^*_{-m}) \ge \hat{\beta})$. Suppose

$$W_m(\alpha) = \begin{cases} \hat{\varepsilon}\alpha & \text{for } \alpha \le \hat{\beta}, \\ K + \hat{\varepsilon}(\alpha - (\hat{\beta}+1)) & \text{for } \alpha > \hat{\beta}, \end{cases}$$

where $\hat{\varepsilon} > 0$ and $K > \hat{\varepsilon}\hat{\beta}$. While $W_m(\alpha)$ is strictly increasing, picking $\hat{\varepsilon}$ sufficiently small allows us to approximate the step function $K \times 1_{\alpha \ge \hat{\beta}+1}$, where $1_{\alpha \ge \hat{\beta}+1}$ is an indicator function that arrivals equal or exceed $\hat{\beta}+1$. Thus, if agent m goes to some bin q, we have that $\mathbb{E}[W_m(\alpha_q(\mathbf{\Pi}_{-m}) + 1)] \approx K \times \Pr(\alpha_q(\mathbf{\Pi}^*_{-m}) \ge \hat{\beta})$. For sufficiently small $\hat{\varepsilon}$, agent m would prefer s to t, contradicting that $\mathbf{\Pi}^*$ is independent of the delay cost function. □

Proof of Lemma 3. Consider an interval completely within bin t, i.e., $[s, s+\Delta) \subset [t, t+1)$ for $\Delta > 0$. Because the bin size is fixed at one, an arrival to t is assigned to the interval $[s, s+\Delta)$ with probability Δ. Let a_Δ denote the number of arrivals in $[s, s+\Delta)$. For all k

$$\begin{aligned}
\Pr(a_\Delta = k) &= \Pr(a_\Delta = k \mid \alpha_t \ge k) \Pr(\alpha_t \ge k) \\
&= \sum_{n=0}^{\infty} \Pr(a_\Delta = k \mid \alpha_t = k+n) \Pr(\alpha_t = k+n) \\
&= \sum_{n=0}^{\infty} \binom{k+n}{k} \frac{\Delta^k (1-\Delta)^n \lambda^{k+n} e^{-\lambda}}{(k+n)!} \\
&= \sum_{n=0}^{\infty} \frac{(\Delta\lambda)^k ((1-\Delta)\lambda)^n e^{-\lambda(\Delta+(1-\Delta))}}{k!n!} \\
&= \frac{(\Delta\lambda)^k e^{-\lambda\Delta}}{k!} \sum_{n=0}^{\infty} \frac{((1-\Delta)\lambda)^n e^{-\lambda(1-\Delta)}}{n!} = \frac{(\Delta\lambda)^k e^{-\lambda\Delta}}{k!}.
\end{aligned}$$

a_Δ is a Poisson random variable. Because arrivals to the bins of a discrete-time Poisson process are independent, the result can be extended to arbitrary intervals.

Now consider two disjoint intervals $[s, s+\Delta)$ and $[s', s'+\Delta')$ for $\Delta, \Delta' > 0$ that both fall completely within bin t, i.e., $s' \geq s+\Delta$ and $[s, s'+\Delta') \subset [t, t+1)$. Let a_Δ and $a_{\Delta'}$ be the respective arrivals in $[s, s+\Delta)$ and $[s', s'+\Delta')$. We show that a_Δ and $a_{\Delta'}$ are independent:

$$\begin{aligned}
&\Pr(a_\Delta = k, a_{\Delta'} = j) \\
&\quad = \sum_{n=0}^{\infty} \Pr(a_\Delta = k, a_{\Delta'} = j \mid \alpha_t = k+j+n) P(\alpha_t = k+j+n) \\
&\quad = \sum_{n=0}^{\infty} \frac{(k+j+n)!}{k!j!n!} \Delta^k (\Delta')^j (1-\Delta-\Delta')^n \frac{\lambda^{k+j+n} e^{-\lambda}}{(k+j+n)!} \\
&\quad = \frac{(\Delta\lambda)^k e^{-\lambda\Delta}}{k!} \frac{(\Delta'\lambda)^j e^{-\lambda\Delta'}}{j!} \sum_{n=0}^{\infty} \frac{((1-\Delta-\Delta')\lambda)^n e^{-\lambda(1-\Delta-\Delta')}}{n!} \\
&\quad = \frac{(\Delta\lambda)^k e^{-\lambda\Delta}}{k!} \frac{(\Delta'\lambda)^j e^{-\lambda\Delta'}}{j!}.
\end{aligned}$$

Because the arrivals in distinct bins are independent, the result easily extends to disjoint intervals of arbitrary lengths. We have thus shown that $\{\hat{A}(t): t \geq 1\}$ has independent increments and that $\hat{A}(t+\Delta) - \hat{A}(t)$ has a Poisson distribution with parameter $\lambda\Delta$. Hence, $\{\hat{A}(t): t \geq 1\}$ is a continuous time Poisson process. □

Proof of Theorem 6. Consider the arrivals α_t^n in system n into bin t, where $1 \leq t \leq T_n$. Given that α_t^n is a binomial random variable it is well known that it converges in distribution to a Poisson random variable with parameter λ (Ross 1983). All bin arrivals converge to identically Poisson distributed random variables. We now establish limiting independence by showing that the joint distribution of any finite collection of bin arrivals tends to the product of their marginal distributions. Consider the arrivals in system n into any two bins $1 \leq s < t \leq T_n$. Their joint distribution is multinomial:

$$\begin{aligned}
&\Pr(\alpha_s^n = k_s, \alpha_t^n = k_t) \\
&\quad = \frac{M_n!}{k_s!k_t!(M_n-k_s-k_t)!} \left(\frac{1}{T_n}\right)^{k_s} \left(\frac{1}{T_n}\right)^{k_t} \left(1-\frac{2}{T_n}\right)^{M_n-k_s-k_t} \\
&\quad = \frac{M_n!}{k_s!k_t!(M_n-k_s-k_t)!} \left(\frac{\lambda}{M_n}\right)^{k_s} \left(\frac{\lambda}{M_n}\right)^{k_t} \left(1-\frac{2\lambda}{M_n}\right)^{M_n-k_s-k_t}.
\end{aligned}$$

Using $M_n/T_n = \lambda$, this is equivalent to

$$\begin{aligned}
&\Pr(\alpha_s^n = k_s, \alpha_t^n = k_t) \\
&\quad = \frac{M_n(M_n-1)\cdots(M_n-(k_s+k_t-1))}{k_s!k_t!} \\
&\qquad \cdot \left(\frac{\lambda}{M_n}\right)^{k_s} \left(\frac{\lambda}{M_n}\right)^{k_t} \left(1-\frac{2\lambda}{M_n}\right)^{M_n-k_s-k_t} \\
&\quad = 1\left(1-\frac{1}{M_n}\right)\cdots\left(1-\frac{k_s+k_t-1}{M_n}\right) \frac{\lambda^{k_s}}{k_s!} \frac{\lambda^{k_t}}{k_t!} \frac{(1-2\lambda/M_n)^{M_n}}{(1-2\lambda/M_n)^{k_s+k_t}}.
\end{aligned}$$

Standard limit arguments show that

$$1\left(1-\frac{1}{M_n}\right)\cdots\left(1-\frac{k_s+k_t-1}{M_n}\right) \to 1,$$

$$\left(1-\frac{2\lambda}{M_n}\right)^{k_s+k_t} \to 1,$$

and

$$\left(1-\frac{2\lambda}{M_n}\right)^{M_n} \to e^{-2\lambda},$$

so that

$$\lim_{n\to\infty} P(\alpha_s^n = k_s, \alpha_t^n = k_t) = \frac{\lambda^{k_s} e^{-\lambda}}{k_s!} \frac{\lambda^{k_t} e^{-\lambda}}{k_t!}.$$

This argument extends to any finite collection of bins. The arrival process thus converges in distribution to a sequence of independent Poisson random variables with rate λ, or a discrete-time Poisson process.

For the second part, consider the $M_n - \hat{A} > 1$ agents who have not arrived by bin $\hat{t}$. Their arrival times are independent and uniformly distributed over the remaining $T_n - \hat{t}$ bins. A representative agent arrives in next k bins with probability $k/(T_n - \hat{t})$. Because agents act independently, we have

$$\begin{aligned}
\Pr(\nu_n \leq k \mid A_n(\hat{t}) = \hat{A}) &= 1 - (1 - k\lambda/(M_n - \lambda\hat{t}))^{M_n-\hat{A}} \\
&= 1 - (1 - k\lambda/Z_n)^{Z_n} (1 - k\lambda/Z_n)^{\lambda\hat{t}-\hat{A}},
\end{aligned}$$

where $Z_n = M_n - \lambda\hat{t}$. Because $Z_n \to \infty$, similar limiting arguments as above establish the desired result. □

References

Cleveland, B., J. Mayben. 1997. *Call Center Management on Fast Forward*. Call Center Press, Annapolis, MD.

Dewan, S., H. Mendelson. 1990. User delay costs and internal pricing for a service facility. *Management Sci.* **36**(12) 1502–1517.

Feller, W. 1957. *An Introduction to Probability Theory and Its Applications: Volume I*, 2nd ed. John Wiley and Sons, New York.

Fudenberg, D., J. Tirole. 1985. Preemption and rent equalization in the adoption of new technology. *Rev. Econom. Stud.* **52** 383–402.

Fudenberg, D., J. Tirole. 1986. A theory of exit in duopoly. *Econometrica* **54** 943–960.

Fudenberg, D., J. Tirole. 1996. *Game Theory*. MIT Press, Cambridge, MA.

Glazer, A., R. Hassin. 1983. ?/*M*/1: On the equilibrium distribution of customer arrivals. *Eur. J. Oper. Res.* **13** 146–150.

Hassin, R. J., M. Haviv. 2003. *To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems*. Kluwer Academic Publishers, Boston, MA.

Larsen, R. J., M. L. Marx. 1986. *An Introduction to Mathematical Statistics and Its Applications*, 2nd ed. Prentice-Hall, Englewood Cliffs, NJ.

Mendelson, H. 1985. Pricing computer services: Queuing effects. *Comm. ACM* **28**(3) 312–321.

Mendelson, H., S. Whang. 1990. Optimal incentive-compatible priority pricing for the $M/M/1$ queue. *Oper. Res.* **38**(5) 870–883.

Naor, P. 1969. The regulation of queue size by levying tolls. *Econometrica* **37**(1) 15–24.

Ostrovsky, M., M. Schwarz. 2003. The adoption of standards under uncertainty. Working paper, Harvard University, Cambridge, MA.

Park, A., L. Smith. 2003. Caller number five: Timing games that morph from one form to another. Working paper, University of Michigan, Ann Arbor, MI.

Rapoport, A., W. E. Stein, J. E. Parco, D. A. Seale. 2003. Strategic play in single-server queues with endogenously determined arrival times. *J. Econom. Behavior Organ.* Forthcoming.

Ross, S. M. 1983. *Stochastic Processes*. John Wiley and Sons, New York.

Seale, D. A., J. E. Parco, W. E. Stein, A. Rapoport. 2003. Joining a queue or staying out: Effects of information structure and service time on arrival and exit decisions. Working paper, University of Arizona, Tucson, AZ.

Shaked, M., J. G. Shanthikumar. 1994. *Stochastic Orders and Their Applications*. Academic Press, San Diego, CA.

Stidham, S. 1992. Pricing and capacity decisions for a service facility: Stability and multiple local optima. *Management Sci.* **38**(8) 1121–1139.

Van Mieghem, J. A. 2000. Price and service discrimination in queueing systems: Incentive-compatibility of Gcμ scheduling. *Management Sci.* **46**(9) 1249–1267.

Vickrey, W. S. 1969. Congestion theory and transportation investment. *Amer. Econom. Rev.* **59**(2) 251–260.

Yechali, U. 1971. On optimal balking rules and toll charges in the $GI/M/1$ queuing process. *Oper. Res.* **19**(2) 349–370.