Invite Your Friend and You'll Move up in Line: Optimal Design of Referral Priority Programs

Luyi Yang
University of California, Berkeley, Haas School of Business, luyiyang@berkeley.edu

This paper studies the optimal design of referral priority programs, in which customers on a waitlist can jump the line by inviting their friends to also join the waitlist. Recent years have witnessed a growing presence of referral priority programs as a novel customer acquisition strategy for firms that maintain a waitlist. Different variations of this scheme are seen in practice, raising the question of what should be the optimal referral priority mechanism. We build an analytical model that integrates queueing theory into a mechanism design framework, where the objective of the firm is to maximize the system throughput, i.e., to accelerate customer acquisition as much as possible. Our analysis shows that the optimal mechanism has one of the following structures: full priority, partial priority, first-in-first-out (FIFO), and strategic delay. A full-priority (partial-priority) scheme enables referring customers to get ahead of all (only some) non-referring ones. A FIFO scheme does not provide any priority-based referral incentive. A strategic-delay scheme grants full priority to referring customers, but artificially inflates the delay of non-referring ones. We show that FIFO is optimal if either the base market size or the referral cost is large. Otherwise, partial priority is optimal if the base market size is above a certain threshold; full priority is optimal at the threshold base market size; strategic delay is optimal if the base market size is below the threshold. We also find that referrals motivate the firm to maintain a larger capacity and therefore, can surprisingly shorten the average delay even though more customers sign up and strategic delay is sometimes inserted. Our paper provides prescriptive guidance for launching the optimal referral priority program and rationalizes common referral schemes seen in practice.

Key words: queueing; referrals; priority; customer acquisition; system throughput; mechanism design

1. Introduction

The referral priority program has gained traction as a customer acquisition strategy in recent years, particularly in the technology sector. The increasing shift of mobile and software services to the cloud often puts a strain on the capacity of technology firms providing these services, and, as a consequence, waitlists emerge as a natural means to buffer customer demand (The Verge 2013). Customers sign up on a waitlist and will not be granted access to the service until it is their turn for account activation. A long wait, however, can be bypassed if a referral priority program is put

in place. In such a program, if a customer wishes to get activated sooner, she can jump the line by inviting her friends to also join the waitlist.

Referral programs in practice vary in how much a referring customer can jump, if at all. For example, Robinhood, a financial-service company providing commission-free stock and cryptocurrency trading (Fortune 2015), presents its referral program as follows: "Interested in priority access? Get early access by referring your friends. The more friends that join, the sooner you will get access." Our conversations with Robinhood confirm that they use a *full-priority* referral scheme, in which a customer, by bringing in one friend, can skip over all the non-referring customers.

However, other companies run their referral programs differently. For instance, x.ai, which offers artificial-intelligence-powered email scheduling services (Financial Times 2018), gives the following account of its referral program: "Hang Tight. We appreciate your patience. Want to jump the line? Share below and move on up." They specify that a customer can move up by two spots for each successful referral. In essence, x.ai's referral program is one of *partial priority*, i.e., a referring customer can overtake some of the non-referring customers but not all of them.

Still, other twists exist. Some companies simply run a first-in-first-out (*FIFO*) waitlist without a referral priority program; examples include Dropbox, a file-hosting service (TechCrunch 2011), and Mailbox, an email inbox-management application (Business Insider 2013). On the other hand, Waitlisted.co, a third-party waitlist-management tool that enables customization of referral priority programs, supports a function to artificially augment the queue size, which could be a means to implement *strategic delay*.

The presence of all these variations begs the question of what should be the optimal referral priority program for customer acquisition. Our conversations with Waitlisted.co suggest that many firms are eager to launch a referral priority program, but often lack guidance on whether they should do it and if so, how. Nevertheless, the academic literature on this topic is tenuous, and to the best of our knowledge, the only existing paper is Yang and Debo (2019), who consider two specific mechanisms, FIFO and full priority, but are silent on the broader mechanism design question.

In contrast to traditional referral reward programs in a non-queueing setting, referral priority programs face several unique design challenges. In the former, customers refer for money, the amount of which can be freely set by the firm. In the latter, customers refer for priority, which cannot be arbitrarily promised without regard to its operational feasibility. For instance, a firm that guarantees an extraordinarily short delay for referring customers may find it difficult to deliver it if too many referring customers qualify. Hence, operational achievability governs how referral incentives should be provided. Moreover, in designing a referral priority program, a subtle trade-off exists between customers' joining and referral incentives. While the launch of a traditional referral reward program reinforces customers' willingness to join (because they can get paid for referrals

on top of obtaining the desired service), the launch of a referral priority program may weaken customers' willingness to join, because referrals can exacerbate congestion, making it less favorable for customers to sign up on the waitlist. To the extent that referred customers cannibalize the demand of base customers who would sign up organically otherwise, the referral incentive can be at odds with the joining incentive, and a delicate balance between the two is warranted.

To find the optimal referral priority program, we build an analytical model that combines mechanism design and queueing theory. The sign-up waitlist is modeled as a single-server Markovian queue with homogeneous delay-sensitive customers. The firm's objective is to acquire as many customers as possible, i.e., to maximize the system throughput, by specifying customers' expected delays as a function of how many converted referrals they have. These expected delays reflect how the firm schedules customers in the sign-up queue. To find viable scheduling policies, we employ the achievable region method (Coffman and Mitrani 1980) to ensure the expected delays specified by the firm are operationally achievable. In our base model, each customer has one potential friend to refer; we later demonstrate the robustness of our results in an extension that accommodates multiple friends. Base customers arrive organically (not through referrals) according to a Poisson process with a rate referred to as the base market size. Given the expected delays posted by the firm, each arriving customer decides on her joining (i.e., sign-up) probability and also how much (costly) referral effort to exert if she joins. The more referral effort one exerts, the more likely a friend is to arrive. Upon arrival, a referred customer goes through the same joining and referral decision process. A referral converts if the friend being referred joins the queue. One key feature of referrals is that they make the sign-up process more bursty because once a base customer joins, a random succession of referred ones follow.

We find that in the first-best problem that ignores customers' referral incentives, the firm should always adopt a work-conserving policy; in particular, when the base market size is large, the non-referral FIFO scheme is the optimal mechanism. Despite the ability to compel customers to refer, the firm in the first best still prefers not to under a large base market size because the system is already congested even without referrals and inviting referrals would inevitably further increase congestion, thus crowding out base customers. The gain of referred customers would not make up for the loss of base ones in this case because it would create a more bursty arrival process that prolongs delay and jeopardizes the system throughput. This result immediately implies that in the second-best problem (which acknowledges customers' referral incentives), the firm should turn off the referral program and run a plain FIFO waitlist when the base market size is large. When it is not too large, though, the first best always encourages some referrals, yet the second best cannot generate any when the referral cost is too prohibitive. Therefore, FIFO is again the optimal mechanism (in fact, the only feasible one) in this case.

Under a small referral cost and an intermediate base market size, it is optimal to employ a partial-priority referral scheme, which achieves the first-best solution. Partial priority stimulates referrals, but also prevents customers from referring too much (so that referrals do not dampen the joining incentive). As the base market size gets smaller, the optimal mechanism injects a stronger referral incentive by allocating more priority to referring customers until the base market size reaches a certain cutoff value, at which point, referring customers receive full priority. When the base market size falls below this cutoff value, then any work-conserving policies fail to generate referrals in the second best because under such light congestion, delay is short anyway and customers lack the incentive to refer. As a remedy, the optimal mechanism resorts to strategic delay—a non-work-conserving policy that gives full priority to referring customers but artificially inflates the delay of those who do not refer—in order to make the priority reserved for those who refer look more attractive. In this manner, the optimal mechanism produces referrals, but still not as many as what would be produced in the first best.

We also explore how transfers can restore efficiency and close the gap between the first best and the second best, and how discriminating between base and referred customers by providing them with different waiting-time quotations can further increase the throughput. Furthermore, we study extensions of multiple friends and endogenous capacity to demonstrate the robustness of our insights. In terms of the capacity implications of referrals, we find that referrals prompt the firm to build more capacity, but in order to incentivize referrals, the firm in the second best may refrain from serving customers as fast as it would in the first best. Interestingly, when capacity is endogenous, the introduction of the referral priority program shortens the average delay of the waitlist despite the fact that more customers join the waitlist as a result of referrals and the fact that customers may be strategically delayed.

In short, our unifying framework rationalizes various referral priority programs observed in practice and provides prescriptive guidance for how to tailor the referral program to a firm's business environment. The remainder of the paper is organized as follows. §2 reviews the relevant literature. §3 formulates the model. §4 begins the analysis by examining the first-best problem. §5 continues the analysis and characterizes the structure of the optimal mechanism in the second-best problem. §6 studies several extensions. §7 concludes the paper and discusses future research directions. All the technical proofs are relegated to Appendix A.

2. Related Literature

Our research bridges the marketing literature on customer referrals, and the operations literature on rational queueing.

The marketing literature on customers referrals focuses on the design of referral reward programs that compensate customers with monetary rewards for successful referrals. The stream of literature

stems from Biyalogorsky et al. (2001), who examine how the firm should jointly set the purchase price and referral reward. They find the referral reward program "pays for performance" and thus mitigates the free-riding problem that would arise if the firm were to attract customers with a low price only. Kornish and Li (2010) design the optimal referral bonuses by capturing referrals' quality-signaling role in addition to its informational role. Xiao et al. (2011) investigate how to provide two-way incentives to both referring and referred customers. Lobel et al. (2017) study the impact of the social network structure on the referral-payment design when firms value referrals but can only compensate conversions. Jing and Xie (2011) contrast the referral reward program with group buying, and find group buying to be more favorable if interpersonal communication is highly efficient. Our model builds on this strand of literature and further contributes to it by studying the referral priority program where customers are incentivized not by pecuniary payment but by priority access on a waitlist. As such, the incentives cannot be specified without regard to their operational achievability.

The operations literature on rational queueing often studies how delay and pricing in the queue impact customer behavior. Hassin and Haviv (2003) and Hassin (2016) present comprehensive reviews of the literature. Early works such as Naor (1969) and Edelson and Hildebrand (1975) examine how to price a service facility to regulate customers' queue-joining incentives. The works on rational customer behavior in priority queues are particularly relevant for our paper. Kleinrock (1967) opens up a body of literature on priority auctions in which customers who bid more receive a strictly higher priority than those who bid less (Lui 1985, Glazer and Hassin 1986, Hassin 1995, Afèche and Mendelson 2004). Another branch of this literature investigates the optimal pricing of priorities, such as Mendelson and Whang (1990), Gavirneni and Kulkarni (2016), Gurvich et al. (2019), Wang et al. (2019), or compares priority pricing with other mechanisms such as line-sitting (Cui et al. 2019) and queue-scalping (Yang et al. 2019). All of these papers focus on work-conserving full-priority policies, in which the firm charges a higher price for higher priority, and customers decide whether to join the queue and if so, self-select into their priority class.

Our paper builds on the growing literature that identifies the optimal priority-pricing structure by integrating mechanism design (Myerson 1981) and the achievable region method (Coffman and Mitrani 1980). The seminal work of Afèche (2013) studies the priority-pricing problem from a revenue-maximizing perspective by allowing for non-work-conserving policies and shows inserting strategic delay may be optimal. Katta and Sethuraman (2005) show that the optimal pricing scheme might pool customers of different types into the same priority class if the virtual delay cost is not monotone increasing. Afèche and Pavlin (2016) find lead-time-dependent customer types can drive the appearance of strategic delay, pooling, or pricing out the middle as the optimal scheduling structure. Yang et al. (2017) consider auction mechanisms for customers to trade their waiting

positions and show that partial pooling is the optimal mechanism for an intermediary in charge of the trading platform. While the above papers employ exact analysis that captures stochasticity of the queue (which is also the approach we take), others resort to fluid approximations to model kidney allocation (Su and Zenios 2006) or deterministic relaxations to characterize the behavior of large-scale systems (Maglaras et al. 2018).

One major difference between priority purchasing and referral priority programs is that in the former, whether one purchases priority does not directly affect how congested the queue is, whereas in the latter, referrals bring new customers to the system, thereby directly changing the congestion level. Moreover, in our framework of referral priority programs, not only can full priority and strategic delay emerge as the optimal scheme (as identified by the literature for priority pricing), but FIFO and partial priority can do as well (different from the literature). Note that partial priority in our setting bears a resemblance to pooling in the literature but the underlying structure is quite different. Partial priority still gives some preferential treatment (i.e., a shorter delay) to referring customers, whereas pooling in those papers do not differentiate at all the delay of different types of customers being pooled.

Unlike the above queueing literature which typically involves mechanisms design problems with "hidden information" (customers privately informed of their type), the incentive problem in our model stems from "hidden action": the firm does not observe and thus cannot contract directly on customers' referral effort; rather, it contracts on the verifiable referral outcome (whether a referred customer joins), which is a probabilistic function of the referral effort. To that end, our paper is related to the principle-agent theory (Ross 1973, Holmström 1979). The unique setting of a referral program introduces an interaction among the agents (i.e., customers): the conversion rate of a referral depends on whether a friend joins, or the expected utility of joining, which, in turn, is a function of how likely a referral converts. Moreover, the queueing setting links a customer' joining decision with those of others. Therefore, the novel features of our problem distinguish it from those in the literature on principle-agent theory.

Finally, to the best of our knowledge, the only other paper that studies referral priority programs is Yang and Debo (2019), who restrict attention to the full-priority referral scheme under fixed capacity. While we build on their model, we also take one step further by endogenizing both the scheduling policies (in the main model) and the capacity (in an extension). In particular, we complement Yang and Debo (2019) by showing that even when the referral program based on full priority backfires or fails to work, the firm may still generate referrals and acquire more customers by launching a carefully designed referral program based on partial priority or strategic delay.

3. Model

We model a sign-up waitlist for a firm's service as a single-server queueing system that has i.i.d. exponential service times with mean $1/\mu$. Parameter μ is the capacity of the firm, the rate at which the firm takes customers (if any) off the waitlist for access to the service. The base model focuses on the case of exogenous capacity; §6.3 extends the analysis to endogenous capacity. *Base customers* arrive to the system according to a Poisson process with rate or *base market size* Λ. These base customers are aware of the service themselves and arrive spontaneously (i.e., not through referrals). When we say a customer arrives, it means that a customer becomes aware of the service and may potentially sign up on the waitlist (i.e., join the queue). An arrival does not necessarily translate to a sign-up, as we shall see next.

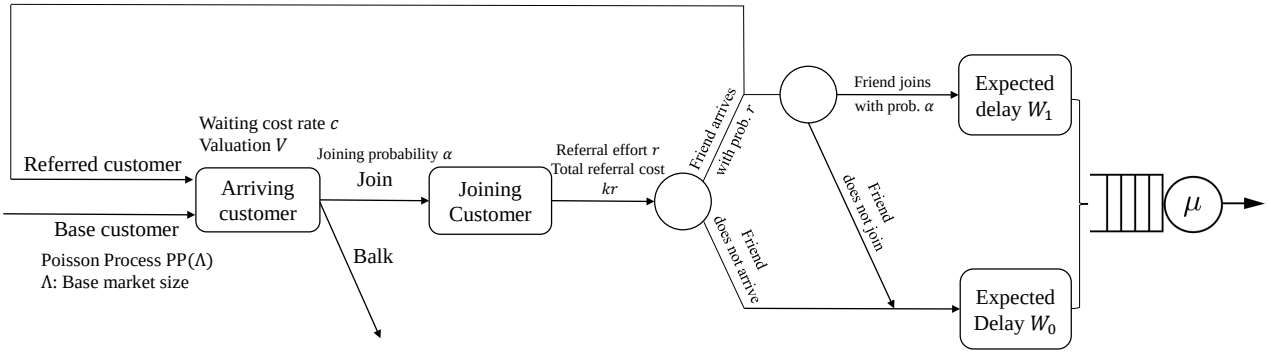
For each arriving customer, her (gross) valuation of the service is V, and her waiting cost per unit time spent in the queueing system (including the time at the server[1]) is c. The waiting cost captures time preference: *ceteris paribus*, customers would rather receive the service earlier than later. Thus, a long wait would diminish a customer's net valuation of the service, thereby reducing her willingness to join the queue. Upon arrival, each customer decides whether to join the queue or balk. If a customer balks, her utility is normalized to zero. If a customer joins, she next determines, upon joining, her effort level $r \in [0,1]$ for making a referral, i.e., for inviting a friend to also join the queue. A customer who exerts effort r incurs total referral cost kr, where k is the cost of effort, or (marginal) referral cost; and with probability r, her friend, a *referred customer* responds, i.e., arrives (instantaneously) upon receiving the referral invitation (but does not necessarily join). If a customer does not exert any effort, i.e., $r = 0$, then her friend will definitely not arrive. At the other extreme, if a customer exerts full effort, i.e., $r = 1$, then her friend will arrive with certainty. The referral effort can be interpreted as the effort expended to reach out to a potentially interested friend and educate the friend on the service provided by the firm. This formulation captures in a stylized way the idea that more referral effort on the part of the referring customer translates to a higher response rate, i.e., the friend she approaches is more likely to be convinced to check out the service (i.e., arrive).

A referred customer has the same valuation and waiting cost primitives as a base customer and also decides on joining and referring. The same process continues for a friend's friend and so on. Once a customer joins, she does not renege. We assume $V > c/\mu$ to rule out the trivial case in which no customers join. All model primitives, V, c, k, Λ, μ, are known to the firm, yet customers only know their valuation and cost parameters V, c, k, but not system parameters Λ or μ.

[1] In our context, if a customer is at the server, it means that she will tentatively be the next one to get off the waitlist and access the service unless she is overtaken.

A referral converts, or is successful, if the invited friend joins the queue. Once a referred customer signs up via the unique referral link she receives from the referring customer, the firm identifies the referring customer and credits her with a successful referral. The firm posts expected delays W_1 for customers who make a successful referral, and expected delay W_0 for those who do not refer or make an unsuccessful referral attempt. Customers do not know the real-time queue length or underlying scheduling policy, and cannot reliably estimate their own delays (because, in part, they are uninformed of system parameters Λ or μ); therefore, they act on the expected delays $\boldsymbol{W} \triangleq (W_1, W_0)$ posted by the firm. A customer's strategy is a pair (α, r), where $\alpha \in [0,1]$ is the joining probability and and $r \in [0,1]$, the referral effort. By allowing α to be either from $\{0,1\}$ or a fractional number between 0 and 1, we accommodate both pure and mixed strategies of joining. Figure 1 illustrate the model setup.

Figure 1 Illustration of the Model



Given the posted expected delays $\boldsymbol{W}$, if all other customers play (α, r), the expected utility for a customer who plays (α', r') is

$$U(\alpha', r' | \alpha, r, \boldsymbol{W}) = \alpha'[V - kr' - c(r'\alpha W_1 + (1 - r'\alpha)W_0)].$$

That is, a focal joining customer's expected utility is her valuation of the service, less the referral cost, less the expected waiting cost, which is calculated by accounting for the likelihood of referral conversion. A referral converts if (1) the invited friend arrives, which occurs with probability r' (given the focal customer's strategy (α', r')), and (2) the friend joins, which occurs with probability α (given all other customers' strategy (α, r), which represents the focal customer's friend's strategy). Therefore, with probability $r'\alpha$ $(1 - r'\alpha)$, the original customer's expected delay is W_1 (W_0). Given $\boldsymbol{W}$, a symmetric Nash equilibrium (α, r) satisfies

$$(\alpha, r) \in \arg\max_{\alpha', r'} U(\alpha', r' | \alpha, r, \boldsymbol{W}). \tag{1}$$

That is, in equilibrium, a customer's joining and referral strategy is the best response to itself. To explicate the equilibrium condition in (1), note that the expected utility $U(\alpha', r'|\alpha, r, \boldsymbol{W})$ is linear in r' with slope $\alpha'[-k + c\alpha(W_0 - W_1)]$. Thus, in a nontrivial equilibrium with $\alpha > 0$, the referral effort r has the following equilibrium properties:

(i) if $-k + c\alpha(W_0 - W_1) > 0$, $r = 1$;

(ii) if $-k + c\alpha(W_0 - W_1) < 0$, $r = 0$;

(iii) if $-k + c\alpha(W_0 - W_1) = 0$, $r \in [0, 1]$.

In Case (i): the marginal benefit of a referral, $c\alpha(W_0 - W_1)$, exceeds the marginal referral cost k, so customers always gain from exerting more referral effort, and thus they end up exerting full effort; in Case (ii), the marginal benefit of a referral is less than the marginal cost, so customers exert zero referral effort; in Case (iii), the marginal benefit of a referral is equal to the marginal cost, so any possible referral effort would give the same expected utility and thus constitute an equilibrium.

Queueing Dynamics. Since, referred customers in our model arrive as soon as they receive an invitation, the joining process of customers is a compound Poisson process. In other words, the queueing system has *batch arrivals*. Each arrival batch is initiated by a base customer, followed by a succession of referred customers who sign up one after another until a customer either stops making a referral or does not refer successfully. For any equilibrium (α, r) that results in a stable queueing system, the effective arrival rate (joining rate) is $\Lambda\alpha$; the batch size N is a random variable following a geometric distribution with

$$\Pr(N = k) = (1 - r\alpha)(r\alpha)^{k-1}, \quad k = 1, 2, \ldots, \qquad \mathbb{E}[N] = \frac{1}{1 - r\alpha}.$$

The *system throughput* is $\Lambda\alpha/(1 - r\alpha)$. Out of a batch of size N, the first $N - 1$ customers in the batch refer successfully, and the last customer does not. Thus, the throughput of customers with one successful referral is $\Lambda r\alpha^2/(1 - r\alpha)$, which is $r\alpha$ fraction of the joining customers, whereas the throughput of non-referring customers (including those who do not refer and those who refer in vain alike) is $\Lambda\alpha$, which is $(1 - r\alpha)$ fraction of the joining customers. A key feature of the batch arrival process (i.e., compound Poisson process) is that it is more *bursty* than the original unit arrival process (i.e., Poisson process). Referrals exacerbate the variability of customer arrivals because in the absence of referrals, the only source of variability is when a customer arrives, but in the presence of referrals, there is also variability in how many referred customers an initial base customer can bring in (i.e., the batch size itself is random), making the customer arrival process even more unpredictable. Such burstiness is at the heart of our subsequent analysis.

Operationally Achievable Delays. Thus far, we have taken the posted expected delays $\boldsymbol{W}$ as given. However, these two quantities cannot be arbitrary, but instead should be *operationally achievable*. Specifically, we require that $\boldsymbol{W}$ agree with the realized average steady-state delays given capacity μ, a particular (nonanticipative and regenerative) scheduling policy, and customers' joining and referral decisions (α, r) induced by $\boldsymbol{W}$. This consistency requirement is standard in the literature (Afèche 2013, Afèche and Pavlin 2016, Yang et al. 2017) and is typically captured by constraints imposed on $\boldsymbol{W}$ ensuring that it must lie within the achievable region (Coffman and Mitrani 1980). The strength of this achievable region method is that it focuses on the scheduling outcome, the expected delays, rather than specific, potentially complex scheduling policies that implement this outcome. We apply this achievable region method. Consistent with the literature (Afèche 2013, Afèche and Pavlin 2016), we allow for non-work-conserving policies for comprehensiveness and preemptive policies for simplicity. Lemma 1 specifies the operationally achievable region for $\boldsymbol{W}$, accounting for the queueing dynamics (e.g., batch arrivals) described above.

LEMMA 1. *For any (α, r) pair that satisfies $\mu > \Lambda\alpha/(1-r\alpha)$, expected delays $\boldsymbol{W} = (W_1, W_0)$ is operationally achievable if and only if*

$$r\alpha W_1 + (1-r\alpha)W_0 \geq \frac{1}{\mu(1-r\alpha)-\Lambda\alpha}, \tag{2a}$$

$$W_1 \geq \frac{1}{\mu(1-r\alpha)-\Lambda r\alpha^2}, \tag{2b}$$

$$W_0 \geq \frac{1}{\mu-\Lambda\alpha}. \tag{2c}$$

The right-hand side of Constraint (2a), $1/(\mu(1-r\alpha)-\Lambda\alpha)$ is the mean delay of a batch-arrival queue with capacity μ, arrival rate $\Lambda\alpha$ and mean batch size $1/(1-r\alpha)$ (see the description of the queueing dynamics). Thus, Constraint (2a) ensures that any scheduling policy cannot reduce the total workload. The right-hand side of Constraint (2b), $1/(\mu(1-r\alpha)-\Lambda r\alpha^2)$, is the mean delay of a batch-arrival queue with capacity μ, arrival rate $\Lambda r\alpha^2$ and mean batch size $1/(1-r\alpha)$. The mean-delay expressions on the right hand sides of Constraints (2a) and (2b) follow from Yang and Debo (2019). The right-hand side of Constraint (2c) is the mean delay of an $M/M/1$ queue with capacity μ and arrival rate $\Lambda\alpha$. Constraints (2b) and (2c) essentially require that the posted expected delay for customers in each group (for both the referring and non-referring groups) cannot be shorter than what they would get if their group were fully prioritized over the other group. Intuitively, Constraints (2a)-(2b) prevent the firm from over-promising delays too short to be achievable.

If Constraint (2a) is binding, then we have work-conserving scheduling policies. If Constraint (2b) is also binding, then $\boldsymbol{W}$ corresponds to a *full-priority* (work-conserving) scheme in which those with a successful referral can jump over all the non-referring customers. If Constraint (2a) is binding but

Constraint (2b) is not, and $W_1 < W_0$, then $\boldsymbol{W}$ implies a *partial-priority* (work-conserving) scheme in which those with a successful referral can jump over some of the non-referring customers but not all of them. If Constraint (2a) is non-binding, then $\boldsymbol{W}$ must be operationalized by a non-work-conserving scheme, which we refer to as *strategic delay*. Finally, if $W_1 = W_0 = 1/(\mu(1-r\alpha) - \Lambda\alpha)$, i.e., a customer who makes a successful referral does not receive any delay reduction, then such $\boldsymbol{W}$ effectively reduces to a *FIFO* scheme.

3.1. The Mechanism Design Formulation

We now formalize the firm's mechanism design problem. The firm's objective is to acquire as many customers as possible. Hence, it aims to maximize the system throughput $\Lambda\alpha/(1-r\alpha)$, the average number of sign-ups per unit time. It chooses expected delays $\boldsymbol{W}$ to induce the optimal joining and referral strategies (α, r) from customers. The firm is subject to the individual rationality (IR) constraint, the incentive compatibility (IC) constraint, the operational achievability (OA) constraints, and the stability constraint, as specified below.

Problem 1.

$$\max_{W_1\geq 0, W_0\geq 0, r\in[0,1], \alpha\in[0,1]} \quad \frac{\Lambda\alpha}{1-r\alpha}, \tag{3}$$

$$\text{s.t.} \quad V - kr - c[r\alpha W_1 + (1-r\alpha)W_0] \geq 0, \tag{4}$$

$$r \in \arg\max_{r'} V - kr' - c[r'\alpha W_1 + (1-r'\alpha)W_0], \tag{5}$$

$$r\alpha W_1 + (1-r\alpha)W_0 \geq \frac{1}{\mu(1-r\alpha) - \Lambda\alpha}, \tag{6a}$$

$$W_1 \geq \frac{1}{\mu(1-r\alpha) - \Lambda r\alpha^2}, \tag{6b}$$

$$W_0 \geq \frac{1}{\mu - \Lambda\alpha}. \tag{6c}$$

$$\frac{\Lambda\alpha}{1-r\alpha} < \mu. \tag{7}$$

The objective of Problem 1 is to maximize the system throughput in (3) through the referral program. Constraint (4) is the IR constraint enforcing that the expected utility from joining should be nonnegative. Constraint (5) is the IC constraint which ensures that choosing effort level r indeed maximizes a joining customer's expected utility, given the expected delays W_1, W_0 and joining probability α of other (referred) customers. To operationalize Constraint (5), see the discussion following the equilibrium condition (1). While the IR and IC constraints explained above specify customer strategies (α, r) for given expected delays $\boldsymbol{W}$, Constraints (6a)-(6c) are OA constraints (as introduced in Lemma 1) that, in turn, guarantee the operational achievability of expected delays $\boldsymbol{W}$ given customer strategies (α, r). Together, Constraints (4) through (6c) close the feedback loop

between the posted delays and customers' joining and referral strategies. Finally, Constraint (7) guarantees the stability of the queueing system.

To avoid triviality, we henceforth impose Assumption 1.

ASSUMPTION 1. $V > c/\mu$.

3.2. Remarks on the Model

Customers in our model are ex-ante homogeneous; their valuation and cost parameters do not vary across individuals. However, referrals create ex-post heterogeneity among customers since referral response and conversion can both be probabilistic. To that end, assuming customers are homogeneous ex-ante presents the simplest setup to cleanly deliver our results.

While we frame the referral strategy r as how much effort to exert, one can alternatively interpret it as the probability of making a referral. As such, any r between 0 and 1 would essentially be a mixed referral strategy (whereas $\alpha \in (0,1)$ represents a mixed joining strategy). That is, if a customer were to refer, her friend would arrive for sure, but a customer may not always refer when $r \in (0,1)$. Both interpretations (referral effort or probability) are valid in our model.

We draw on Yang and Debo (2019) (which fixes the scheduling policy to full priority) for some key modeling choices of our model, one of which is the implicit assumption that each customer has a single friend she can potentially refer. The single-referral assumption is in line with the marketing literature on customer referrals (e.g., Biyalogorsky et al. 2001, Kornish and Li 2010), and, due to variations in referral outcomes, a single referral can effectively create two priority classes ex post, a common focus of the operations literature on priority queues (e.g., Afèche 2013, Gavirneni and Kulkarni 2016). In §6.1, we demonstrate the robustness of our results in a model of multiple friends.

4. The First-Best Problem

In this section, we study the first-best problem as if customers would exert exactly the amount of referral effort dictated by the firm. Technically, the first-best problem relaxes IC constraint (5) in Problem 1. Proposition 1 characterizes the optimal solution to the first-best problem.

PROPOSITION 1. *The first-best problem solution is always work-conserving. In the first best,*

(i) if $\Lambda < \mu - c/V$, it is optimal to generate referrals; the joining probability α^{FB} and referral effort r^{FB} are

$$\alpha^{FB} = 1, \quad r^{FB} = \frac{-\sqrt{4k\mu(c+\Lambda V-\mu V)+(-k\Lambda+k\mu+\mu V)^2}-k\Lambda+k\mu+\mu V}{2k\mu} > 0;$$

(ii) if $\Lambda \geq \mu - c/V$, it is optimal to generate no referrals; the joining probability α^{FB} and referral effort r^{FB} are

$$\alpha^{FB} = \frac{\mu - c/V}{\Lambda}, \quad r^{FB} = 0.$$

In the first best, the firm follows work-conserving policies; in other words, Constraint (6a) should be binding. Note that aside from work conservation, the first best does not specify values of $\boldsymbol{W}$, which implies that any admissible scheduling policy (which would induce operationally achievable $\boldsymbol{W}$) is optimal in the first best. This is because the first best can dictate the referral effort without worrying about customers' referral incentives. In our model with homogeneous customers, given (α, r), any work-conserving scheduling policy yields the same mean delay, $1/(\mu(1-r\alpha)-\Lambda\alpha)$, and thus the IR constraints (joining incentives) would not be affected by the change of scheduling policies.

When the base market size is small ($\Lambda < \mu - c/V$), the system throughput without referrals is small, and it is intuitive that the firm in the first best wants to generate referrals ($r^{\mathrm{FB}} > 0$) to increase system throughput. Moreover, it can be shown (see the proof of Proposition 1) that r^{FB} is decreasing in Λ, suggesting that a firm facing a smaller base market size would encourage a higher referral intensity. Note that in this case of relatively light congestion, customers' joining probability without referrals is one because the waiting cost, $c/(\mu - \Lambda)$ is less than service value V, leaving customers with positive surplus. Hence, introducing referrals can benefit the system throughput without affecting the joining probability (which is still equal to one). The optimal first-best referral strategy adds just enough congestion to the system that customers no longer keep any positive surplus but are still willing to join the queue with certainty.

Nevertheless, when the base market size is large ($\Lambda \geq \mu - c/V$), the first best prescribes a shutdown of the referral program. This is an intriguing result given that the firm in the first best can always compel customers to refer without confronting any pushback. Also note that the cutoff value $\mu - c/V$ does not depend on cost of effort k, so even when referrals are free, the firm would still forbid them in the first best. The crux is that the firm must provide enough joining incentives and ensure individual rationality. When the base market size is large, the system is already somewhat congested without referrals (such that only a fraction of customers join, manifested by joining probability $\alpha^{\mathrm{FB}} < 1$), introducing referred customers (making r positive) to the system would further increase congestion, diminishing customers' joining probability (decreasing α), and, in particular, turning away some base customers who would otherwise join. However, when referred customers cannibalize the demand of base customers, it does not benefit the overall system throughput, and therefore, the firm would prefer banning referrals altogether. The burstiness of the batch-arrival process driven by referrals plays a crucial role here. Burstiness would prolong delay, making the loss of base customers outweigh the addition of referred customers in this case, and the overall system throughput would suffer. In sum, Proposition 1 shows that the first-best problem welcomes referrals only to the extent that they do not crowd out base customers.

5. The Second-Best Problem

In this section, we study the second-best problem by bringing back IC constraint (5) to Problem 1. Recall that a unique feature of the first best is that it only specifies joining probability α and referral effort r but does not uniquely pin down expected delays (W_1, W_0). We will show in this section when and how the first-best solution can be achieved in the second-best problem by properly setting (W_1, W_0) so as to satisfy the IC constraints. We will also show the structure of the scheduling policy in terms of (W_1, W_0) when the first best cannot be achieved.

5.1. The Optimal Mechanism

When the market size is large, i.e., $\Lambda > \mu - c/V$, the first best requires that no referrals be generated. This outcome can be readily achieved in the second best by shutting down the referral priority program, i.e, setting $W_1 = W_0$ to give zero referral incentives. Hence, the firm should run a plain FIFO waitlist. This result is summarized in Proposition 2.

PROPOSITION 2. *If $\Lambda \geq \mu - c/V$, the first best can be achieved by a non-referral FIFO mechanism.*

The following Proposition 3 pins down the second-best problem under a large referral cost.

PROPOSITION 3. *If $k \geq V - c/\mu$, there does not exist any (second-best) mechanism that can generate referrals. Hence, non-referral FIFO is the only equilibrium.*

When the referral cost is large, i.e., $k \geq V - c/\mu$, then generating referrals would be infeasible in the second-best problem, let alone optimal. To see why this result holds, first consider the case of full referral effort $r = 1$. This case cannot arise under such a large referral cost because it would violate the IR constraint ($V - k - c/\mu \leq 0$ and customers' expected delay must be larger than c/μ). Next, consider partial referral effort $r \in (0, 1)$. If we use the mixed-referral-strategy interpretation, customers in this case must be indifferent to referrals, which implies their expected utility would be equal to $V - cW_0$, what they would get should they choose not to refer. Moreover, the marginal cost of a referral should be equal to its marginal benefit, i.e., $k = c\alpha(W_0 - W_1) \leq cW_0 - cW_1$. Thus, $V - cW_0 \leq V - k - cW_1$; since $cW_1 > c/\mu$ for $r > 0$, the IR constraint is again violated under a large enough referral cost $k \geq V - c/\mu$. Intuitively, Proposition 3 suggests that if the cost of referrals is too prohibitive, then there is no way for the benefit of delay cost reduction to cover the referral cost. Therefore, no referrals can be generated; the system is again back to FIFO.

Combining Propositions 2 and 3 yields two scenarios: (1) if the base market size is large, i.e., $\Lambda \geq \mu - c/V$, then regardless of the referral cost, the first best can indeed be achieved in the second best by non-referral FIFO; (2) when the base market size is small and the referral cost is large, i.e., $\Lambda < \mu - c/V$ and $k \geq V - c/\mu$, then the second best delivers the non-referral FIFO scheme, but

it falls short of the throughput in the first best which would instead require referrals be generated to acquire more customers than a FIFO scheme would. Propositions 2 and 3 together rationalize the use of FIFO under either a large base market size or a large referral cost.

Next, we characterize the optimal mechanism when neither the base market size nor the referral cost is too large, i.e., $\Lambda < \mu - c/V$ and $k < V - c/\mu$.

PROPOSITION 4. *If $k < V - c/\mu$, then there exists a unique cutoff value $\Lambda^* \in (0, \mu - c/V)$ such that*

(i) if $\Lambda \in (\Lambda^*, \mu - c/V)$, the first best can be achieved by partial priority with the following expected delays (W_1, W_0):

$$W_1 = \frac{1}{\mu(1-r^{FB}) - \Lambda} - \frac{k(1-r^{FB})}{c}, \quad W_0 = \frac{1}{\mu(1-r^{FB}) - \Lambda} + \frac{kr^{FB}}{c};$$

(ii) if $\Lambda = \Lambda^*$, the first best can be achieved by full priority;

(iii) if $\Lambda < \Lambda^*$, the first best cannot be achieved; strategic delay is optimal in the second best; the joining probability α^{SB}, referral effort r^{SB}, and expected delays (W_1, W_0) are

$$\alpha^{SB} = 1, \quad r^{SB} = \frac{\mu(V-k) - c}{(V-k)(\mu+\Lambda)}, \quad W_1 = \frac{1}{\mu(1-r^{SB}) - \Lambda r^{SB}}, \quad W_0 = W_1 + k/c.$$

Proposition 4-(i) shows that when the base market size is intermediate, i.e., $\Lambda \in (\Lambda^*, \mu - c/V)$, and the referral cost is small, i.e., $k < V - c/\mu$, a partial-priority scheme should be adopted to achieve the first best. We give the closed-form expression of the cutoff value Λ^* in the proof of Proposition 4 in Appendix A. Recall from Proposition 1 that under such an intermediate base market size, the first best would generate referrals to the extent that the extra congestion created does not cannibalize the demand of base customers. As a consequence, a partial-priority scheme is in order because it provides customers with a bit of an incentive to refer (they can jump over some non-referring customers if their referral is successful), but not too strong of an incentive (a customer with a successful referral cannot overtake all of the non-referring customers) to let a crowd-out effect kick in. Partial priority is implementable in this case because the referral cost is relatively small, which ensures that the amount of delay-cost reduction offered by the partial-priority scheme can be sufficient to spur voluntary customer referrals. Therefore, a carefully calibrated partial-priority referral scheme can recover the first-best outcome in equilibrium.

Recall from Proposition 1 that the firm in the first best would encourage more referrals with a smaller base market size (r^{FB} is decreasing in Λ), which must be accomplished in the second best by providing a larger referral incentive, i.e., rewarding customers who refer successfully with a greater amount of priority. As the base market size keeps decreasing, it will eventually hit a critical point at which the optimal mechanism must offer full priority to achieve the first best. That

critical point is $\Lambda = \Lambda^*$, as indicated by Proposition 4-(ii). Based on this result, the applicability of a full-priority referral scheme may seem restrictive because it is optimal only under knife-edge events. However, full priority may still hold some practical appeal because it spares the firm the onus to parameterize the referral priority program, which would be necessary should the optimal mechanism (for instance, partial priority) be adopted. To the extent that the full-priority referral scheme improves the system throughout, the firm may find it good enough and thus prefer this simple, parameter-free mechanism.

Proposition 4-(iii) shows that when the base market size is small, i.e., $\Lambda < \Lambda^*$, the firm in the second best should insert strategic delay to the queue in order to motivate referrals. This second-best solution will not achieve the first best because the latter requires work conservation. However, the second best does preserve one feature from the first best: all arriving customers join the queue (i.e., $\alpha = 1$), which shows once again that the optimal referral program should attract referred customers without losing base customers. The distinction between the first best and second best consists in the lack of sufficient referral effort induced by the second best as compared to that imposed by the first best, i.e., $r^{SB} < r^{FB}$. Hence, the firm implementing the optimal mechanism will still not lure as many referred customers in the second best as it would hope for.

A small base market size implies light congestion in the queue, and even customers who are not entitled to any priority can get served quickly. Therefore, any work-conserving scheduling policy will simply not create enough of a delay differential to incentivize referrals. Thus, the firm must resort to strategic delay to acquire more customers. Specifically, the delay of non-referring customers should be artificially inflated to make the delay for those who refer successfully look shorter and more enticing by comparison. While the idea of intensifying competition (in our context, customers compete to access the service) by making a resource look more scarce than it actually is sounds a reasonable tactic, a common misconception about deploying such a tactic is that the firm should decrease capacity. It turns out that doing so is suboptimal, as indicated by Proposition 4-(iii), which shows that the optimal mechanism only elongates the expected delay for non-referring customers, W_0, while keeping the expected delay for those who refer successfully, W_1, as short as possible by giving them full priority (the OA constraint (6a) is non-binding but (6b) is binding). By contrast, decreasing capacity would increase the expected delays (both W_1 and W_0) across the board. Since the goal of strategic delay is to create a large enough delay differential to incentivize referrals, the firm should slow down in a discriminatory fashion instead of just adopting a lower capacity. Still, one caveat for strategic delay to work is the referral cost not being too large. Otherwise, the firm would face an insurmountable dilemma of creating a delay differential large enough to motivate referrals while keeping the amount of delay small enough to ensure joining. In sum, Proposition 4 rationalizes the use of partial priority, full priorities and strategic delay.

The following Proposition 5 characterizes how the referral cost impacts the cutoff value Λ^*.

PROPOSITION 5. *The cutoff value on the base market size Λ^* is increasing in referral cost k, i.e., $\partial\Lambda^*/\partial k > 0$.*

Proposition 5 implies that with a larger referral cost, the firm will resort to strategic delay (partial priority) more often (less often), i.e., for a wider range of base market sizes (for a narrower range of base market sizes). If the referral cost is very small, the firm can easily motivate referrals without violating work conservation, and therefore, strategic delay would be rarely used. If the referral cost gets relatively large, approaching $V - c/\mu$, then generating referrals would be progressively difficult, and the firm would almost invariably turn to strategic delay. Proposition 5 also implies that for a given base market size that is not too large ($\Lambda < \mu - c/V$), the firm should adopt strategic delay when the referral cost is relatively large (but still smaller than $V - c/\mu$), and should use partial priority when the referral cost is small.

5.2. Numerical Illustrations

To further explain our results, we present three numerical examples that compare the optimal mechanism, the FIFO scheme, and the full-priority referral scheme. The equilibrium conditions for the full-priority case are provided in Appendix B.

EXAMPLE 1. $\mu = 1, c = 1, V = 10, k = 2, \Lambda = 0.85$. See Table 1.

Table 1 The Comparison of Different Schemes in Example 1

	Throughput	r	α	W_1	W_0
FIFO	0.85	0	1	6.667	6.667
Full Priority	0.763	1	0.473	2.968	12.516
Optimal (Partial Priority)	0.8938	0.049	1	8	10

The optimal mechanism is partial priority; it achieves the first best. The full-priority referral scheme leads to a system throughput even lower than FIFO.

In Example 1, under the optimal mechanism, customers expend only a slight referral effort ($r = 0.049$); and all arriving customers join ($\alpha = 1$), as in the FIFO case. Referrals strictly improve the system throughput over FIFO. The optimal mechanism is a partial priority scheme as customers who refer successfully enjoy a modest delay reduction relative to their non-referring counterparts ($W_1 = 8$ versus $W_0 = 10$). By contrast, the improvement in delay is more drastic in the full priority scheme if one makes a successful referral ($W_1 = 2.968$ versus $W_0 = 12.516$). Such a sharp difference in expected delays prompts a strong referral incentive: customers in the full-priority scheme exert the maximum referral effort ($r = 1$). Nevertheless, the strong referral incentive undermines the joining incentive: only a fraction of arriving customers choose to join ($\alpha = 0.473 < 1$). This leads to an unintended consequence: the full-priority scheme in this example is not only suboptimal, but even falls short of the FIFO throughput.

It has been previously identified in Yang and Debo (2019) that a referral program which grants full priority to customers with a successful referral can backfire in the sense that launching such a program may harm the system throughput. Since Yang and Debo (2019) only compare the full-priority scheme with the non-referral FIFO case, their recommendation in this situation is simply to forsake referrals altogether. By contrast, our paper takes one step further by identifying a carefully calibrated partial-priority referral scheme—as a middle ground between FIFO and full priority—that can improve the FIFO system throughput and achieve the first best, despite the undesirable performance of the full-priority referral scheme.

EXAMPLE 2. $\mu = 1, c = 1, V = 10, k = 2, \Lambda = 0.7$. See Table 2.

Table 2 The Comparison of Different Schemes in Example 2

	Throughput	r	α	W_1	W_0
FIFO	0.7	0	1	3.33	3.33
Full Priority	0.742	1	0.515	3.335	12.948
Optimal (Partial Priority)	0.871	0.196	1	8	10

The optimal mechanism is partial priority; it achieves the first best. The full-priority referral scheme leads to a system throughput higher than FIFO.

However, the full priority scheme may also improve the FIFO throughput under other circumstances. Example 2 is one such example. Similar to Example 1, the optimal mechanism in Example 2 also calls for a partial-priority scheme which achieves the first best, but different from Example 1, both the (optimal) partial-priority and full-priority schemes induce a higher system throughput than FIFO in Example 2. The full-priority scheme here acquires referred customers also at the expense of some base customers (i.e., $\alpha < 1$), but the gain in referred customers outweighs the loss of base customers, so the overall throughput is still enhanced. By comparison, the optimal partial-priority scheme seizes referred customers without sacrificing base customers, and the resulting system throughput is even higher.

EXAMPLE 3. $\mu = 1, c = 1, V = 10, k = 2, \Lambda = 0.01$. See Table 3.

Table 3 The Comparison of Different Schemes in Example 3

	Throughput	r	α	W_1	W_0
First-Best	0.0763	0.869	1	-	-
FIFO	0.01	0	1	1.01	1.01
Full Priority	0.01	0	1	1.01	1.01
Optimal (Strategic Delay)	0.0748	0.866	1	8	10

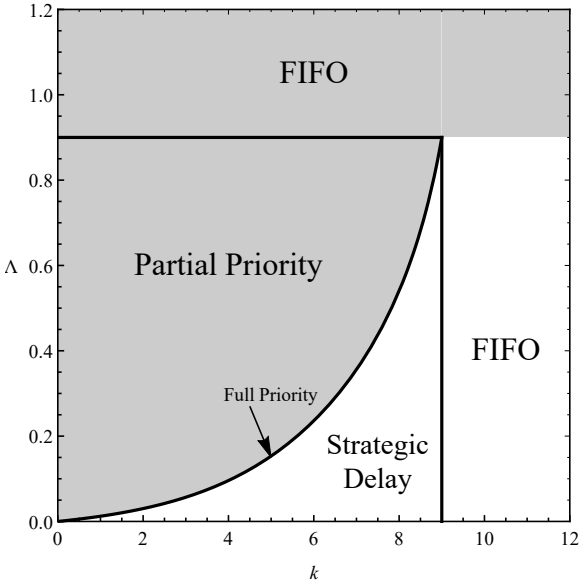
The optimal mechanism is strategic delay. It does not achieve the first best but results in a higher throughput than FIFO. The full-priority scheme cannot generate referrals and thus reduces to FIFO.

In Example 3, the optimal mechanism of strategic delay exhibits a huge throughput improvement over FIFO (by a factor of 7). It still falls short of the first best, albeit only marginally in this case. By contrast, the full-priority scheme cannot stimulate referrals and results in the same equilibrium outcomes as FIFO. Note that the full-priority referral scheme's failure to generate referrals when the base market size is too small has also been established in Yang and Debo (2019). However, instead of concluding referrals are out of the question, we show that under those circumstances, a carefully-designed referral program that involves strategic delay can still "save the day."

5.3. Summary

Based on the analytical results developed in this section, we summarize the structure of the optimal mechanism (illustrated by Figure 2). The firm should turn off any referral program and run a FIFO queue if the base market size is large (Proposition 2) or if the referral cost is large (Proposition 3); in the former case, the first best is achieved, but not necessarily in the latter. If the referral cost is small and the base market size is intermediate, a partial-priority referral scheme is optimal (Proposition 4) and achieves the first best. If both the referral cost and the base market size are small, then the firm should insert strategic delay to the queue to motivate referrals, but it does not achieve the first best (Proposition 4). At the boundary base market size that demarcates the previous two cases, a full-priority referral scheme is the most desirable (Proposition 4). Moreover, this boundary base market size is increasing in the referral cost (Proposition 5), suggesting that strategic delay should be more widely adopted as the optimal mechanism and partial priority less so as the referral cost increases.

Figure 2 The Structure of the Optimal Mechanism



Note. $\mu = 1$, $c = 1$, $V = 10$. In the shaded (unshaded) area, the second best achieves (does not achieve) the first best.

Note that the base market size measures the popularity of the service within the organic customer base in the absence of the referral program. When we comment on the magnitude of the base market size, we are essentially evaluating it against capacity. For example, some web-based services may not be particularly capacity-constrained, and thus the organic demand can be regarded as low by comparison; as a consequence, strategic delay can be employed to boost referrals and jump-start customer acquisition. Some other services may be heavily cloud-based and not easily scalable initially, then the base market size may loom large relative to capacity.

As partial priority and strategic delay can be optimal under certain circumstances, the implementation issues come to the fore. Coffman and Mitrani (1980) and Afèche (2013) propose various approaches to implementing partial priority and strategic delay, respectively. Coffman and Mitrani (1980) primarily comment on two approaches to partial priority, one being probabilistic priority assignment and the other being alternate priorities. In our context, the first approach would assign a referring customer to a priority class only with a certain probability. Thus, even with a successful referral, there is no guarantee of jumping ahead of all non-referring customers. The second approach would shift back and forth between full priority assigned to referring customers and no priority assigned. The period of priority lasts t_1 amount of time, before switching to a period of no priority, which lasts t_2 amount of time, before switching back to a t_1-period of priority, and so on. In practice, however, one commonly used approach to implementing partial priority, as exemplified by the referral program adopted by x.ai, is to let a referring customer move up by a fixed number of spots in the queue. Note that probabilities and time durations are continuous and can be calibrated with ease, making the two approaches proposed by Coffman and Mitrani (1980) flexible enough to match any partial-priority outcome (as specified by the expected delays). The x.ai approach, on the other hand, suffers from the limitation that the number of people to skip over is discrete, and thus such a mechanism cannot always exactly match a given partial-priority outcome.

Afèche (2013) provides a detailed discussion of how strategic delay can be implemented by "idling the server before, reducing its speed during, and delaying the delivery after processing." In our model, if a customer is at the server, it means that she will tentatively be the next candidate to get off the waitlist and access the service. Thus, one approach to implementing strategic delay is to postpone activating the account of any non-referring customer at the top of the waitlist even if the firm is ready to take her off the list. We emphasize that since the optimal strategic-delay policy does not slow down referring customers, they should jump over all the non-referring customers on the list and be taken off the list as soon as possible. Another approach, which is in the spirit of Waitlisted.co's current practice, is to insert a fixed number of artificial customers to the non-priority class, or the class of non-referring customers. Doing so augments the size of the non-priority class and the expected delay therein. Whenever an artificial customer advances to the server, the

firm allocates a certain amount of (random) service time to her just as if she were a real customer, then removes her from the top of the waitlist, and immediately tacks her back onto the bottom of the waitlist. This ensures that there is always a fixed number of artificial customers perpetually circulating in the non-priority class of the queue. Similar to the x.ai approach, the discrete nature of this approach may prevent it from perfectly matching a given strategic-delay scheme.

6. Extensions

This section studies four extensions, namely, multiple friends, achieving the first best through transfers, endogenous capacity, and discriminating between base and referred customers. Each extension relaxes one assumption of the base model at a time while keeping the rest unchanged.

6.1. Multiple Friends

In the base model, each customer has only one friend they can potentially invite; in this subsection, we extend our analysis to multiple friends. Assume that each customer has D friends, where $D \in \mathbb{N}_+$. A joining customer expends effort r_i on referring friend i of hers, $i = 1, \ldots, D$. Denote $\boldsymbol{r} \triangleq (r_1, r_2, \ldots, r_D)$. Hence, the possible numbers of successful referrals are $0, 1, \ldots, D$. Let W_i be the expected delay posted by the firm for a customer who successfully refers i friends, $i = 0, 1, \ldots, D$. Denote $\boldsymbol{W} = (W_0, W_1, \ldots, W_D)$. The firm's objective is to maximize the system throughput by setting $\boldsymbol{W}$ that induces the optimal referral effort $\boldsymbol{r}$ and joining probability α from customers. Note that our base model of a single friend is a special case of this more general model of D friends: This D-friend model reduces to the single-friend model if we set $W_1 = W_2 = \ldots = W_D$, which will induce $r_2 = r_3 = \ldots = r_D = 0$. Given customer strategy $(\boldsymbol{r}, \alpha)$, let $X(\boldsymbol{r}, \alpha)$ denote the number of successful referrals. Then $X(\boldsymbol{r}, \alpha)$ is a random variable following the *Poisson binomial distribution* (Wang 1993, Chen et al. 1994, Harremoes 2001) with parameters $(r_1\alpha, r_2\alpha, \ldots, r_D\alpha)$. The mean of $X(\boldsymbol{r}, \alpha)$ is $m(\boldsymbol{r}, \alpha) \triangleq \alpha \sum_{i=1}^{D} r_i$. The variance of $X(\boldsymbol{r}, \alpha)$ is $\sigma^2(\boldsymbol{r}, \alpha) \triangleq \sum_{i=1}^{D} r_i\alpha(1 - r_i\alpha)$. The probability mass function of $X(\boldsymbol{r}, \alpha)$ is

$$f_i(\boldsymbol{r}, \alpha) \triangleq \Pr(X(\boldsymbol{r}, \alpha) = i) = \sum_{\mathcal{A} \in \mathcal{G}_i} \prod_{k \in \mathcal{A}} (r_k\alpha) \prod_{j \in \mathcal{A}^c} (1 - r_j\alpha),$$

where $\mathcal{G}_i$ is the set of all subsets of i integers that can be selected from $\{1, 2, \ldots, D\}$ and $\mathcal{A}^c = \{1, 2, \ldots, D\} \setminus \mathcal{A}$.

Queueing Dynamics. As in the single-friend model, referrals lead to *batch arrivals* in the D-friend model. Initiated by the arrival of a base customer, each batch itself is generated according to a Galton-Watson branching process (Athreya and Ney 1972) whose progeny distribution is Poisson binomial, i.e., the number of progenies (successful referrals) each individual (joining customer) has is $X(r, \alpha)$. The total batch size is the total number of individuals ever born in the branching

process before the population becomes extinct. Thus, by the standard theory of branching processes, the expected number of the batch size is $1/(1-m(\boldsymbol{r},\alpha))$, and thus, the system throughput is $\Lambda\alpha/(1-m(\boldsymbol{r},\alpha))$. Next, we formulate the firm's first-best problem.

PROBLEM 2.

$$\max_{\boldsymbol{W}\geq 0,\boldsymbol{r}\in[0,1]^D,\alpha\in[0,1]} \quad \frac{\Lambda\alpha}{1-m(\boldsymbol{r},\alpha)} \tag{8}$$

$$\text{s.t.} \quad V - k\sum_{i=1}^{D} r_i - c\sum_{i=0}^{D} f_i(\boldsymbol{r},\alpha)W_i \geq 0, \tag{9}$$

$$\sum_{i=0}^{D} f_i(\boldsymbol{r},\alpha)W_i \geq \frac{\sigma^2(\boldsymbol{r},\alpha)+(2-m(\boldsymbol{r},\alpha))(1-m(\boldsymbol{r},\alpha))}{2(1-m(\boldsymbol{r},\alpha))[\mu(1-m(\boldsymbol{r},\alpha))-\Lambda\alpha]} \tag{10}$$

$$\frac{\Lambda\alpha}{1-m(\boldsymbol{r},\alpha)} < \mu. \tag{11}$$

The objective of Problem 2 is to maximize the system throughput in (8). Constraint (9) is the IR constraint. Constraint (10) is the OA constraint that limits how small the average of the posted expected delays can be. The right-hand side of Constraint (10) follows from Lemma E.1 of Yang and Debo (2019). Constraint (11) is the stability constraint. Note that since $\boldsymbol{W}$ is a $(D+1)$-dimensional vector, in principle, we need $2^{D+1}-1$ inequalities (exponential in D) to specify the operationally achievable region for $\boldsymbol{W}$ (see Coffman and Mitrani 1980). However, for the purpose of the first-best problem, Constraint (10) suffices because the summation $\sum_{i=0}^{D} f_i(\boldsymbol{r},\alpha)W_i$ shows up as a whole in the IR constraint. Proposition 6 below characterizes the first-best problem.

PROPOSITION 6. *The first-best problem for $D>1$ yields the same solution as that for $D=1$.*

Proposition 6 reveals a somewhat surprising result that even if customers have multiple friends, the firm in the first-best world would still only want them to refer at most one friend. That is, the first best of Problem 2 coincides with that of Problem 1 despite the fact that the former is a generalization of the latter. Why does having more referrals from a single customer fail to accelerate customer acquisition (i.e., improve the system throughput)? The crux again lies in how referrals impact the burstiness of the arrival process. If a referring customer reaches out to multiple friends (i.e., $r_i>0$ for multiple i), then it is not as desirable from the firm's perspective as if she exerts all the referral efforts combined on referring one friend, because the former will lead to a more bursty arrival process, which lengthens the delay, making the queue less appealing for customers to join. Such burstiness ultimately works against customer acquisition. Intuitively, if one spreads out her referral efforts across friends, the referral outcome (in terms of the number of referral successes) will be more uncertain (which adds burstiness to the queue) than if one "puts all the eggs in one basket" by focusing on one friend.

While Proposition 6 characterizes the first-best problem, it has significant implications for the second-best problem: whenever the first best can be achieved by the second best in the single-friend model (correspond to the shaded area of Figure 2), then it can be achieved through exactly the same mechanism in the multi-friend model. Propositions 2 and 4, which establish the optimality of FIFO and partial/full priority under their respective conditions, readily carry over. Proposition 6 also implies that whenever the first best cannot be achieved in the single-friend model (corresponding to the unshaded area of Figure 2), neither can it be in the multi-friend model. This is because any mechanism that achieves the first best in the multi-friend model must match its first-best outcome of each customer referring at most one friend; letting customers have more friends will not help close the gap between the first best and second best. Overall, our insights from the base model of a single friend largely extend to the more general case of multiple friends.

6.2. Achieving the First Best Through Transfers

Our base model implicitly focuses on a setting that prohibits transfers between the firm and customers. This is a very reasonable assumption for the applications we model. Many of these waitlists offer products that are "freemiums" in nature and do not charge customers any fees when they sign up. For example, zero-commission trading is Robinhood's primary value proposition, which would be ruined if customers were asked to pay. Further, partially because it is often free to sign up, many waitlists employ a pure priority-based referral scheme without monetary referral incentives. However, to the extent that transfers are permitted, we can fix the inefficiency of the second best and achieve the first best. Proposition 7 below proposes one such scheme.

PROPOSITION 7. *The first best* (α^{FB}, r^{FB}) *defined in Proposition 1, when not achieved in the base model (which prohibits transfers), can be achieved by rewarding a successful referral with full priority plus a cash bonus* P*, and charging each joining customer an admission fee* $r^{FB}P$*, where*

$$P = k - c\frac{\Lambda}{[\mu(1-r^{FB})-\Lambda][\mu(1-r^{FB})-\Lambda r^{FB}]} > 0.$$

In Proposition 7, our proposed mechanism to achieve the first best taxes joining by charging an admission fee and subsidizes referrals by offering a cash bonus in addition to the operational incentive of full-priority access. Moreover, the expected referral bonus each customer receives is equal to $r^{FB}P$, the admission fee each customer pays, which implies that for the firm, the total amount of admission fees collected offsets the total amount of referral bonuses paid out in the long run. Thus, the mechanism is budget-balanced. The cash bonus compensates for the insufficiency of operational incentives and prompts referrals from customers who would otherwise not refer. The admission fee funds the cash bonus, ensuring the firm does not operate at a loss for the sake of growth. Note that mechanisms to achieve the first best through the integration of transfers and

priority access are not unique. An alternative mechanism, for instance, can combine a larger referral bonus with a smaller operational incentive (e.g., partial priority), but the full-priority scheme proposed in Proposition 7 makes the most of the operational leverage of a waitlist; also, keeping the transactional amount to the minimum may be palatable to both the firm and customers.

6.3. Endogenous Capacity

Our base model assumes an exogenous capacity constraint; this subsection endogenizes capacity. We assume that the cost of maintaining capacity μ is $\omega\mu^2/2$ per unit time, where parameter ω measures the ease with which the firm is able to scale capacity (e.g., by purchasing more cloud services). The firm determines capacity μ and the referral mechanism to maximize the objective function $\Lambda\alpha/(1-r\alpha)-\omega\mu^2/2$, i.e., the gain from the system throughput (with the gain from each unit of the throughput normalized to one) less the capacity cost. Since capacity μ is now endogenous, we drop Assumption 1 (which involves μ) and, for ease of exposition, replace it with $V > c/\Lambda$, i.e., customers' valuation of the service is sufficiently high. Let $\bar{\omega} \triangleq \Lambda V^2/\{[\Lambda V^2 + c(V+k)](\Lambda + c/V)\}$. Let μ^{FIFO} denote the firm's optimal capacity choice in a FIFO queue without referrals. Proposition 8 below characterizes the firm's optimal capacity in the first best.

PROPOSITION 8. *In the first best, the optimal capacity μ^{FB} is*

$$\mu^{FB} = \begin{cases} \mu_0, & \text{if } w < \bar{\omega}; \\ \Lambda + c/V, & \text{if } w \in [\bar{\omega}, \frac{V}{\Lambda V + c}); \\ 1/\omega, & \text{if } w \in [\frac{V}{\Lambda V + c}, \frac{V}{2c}); \\ 0, & \text{if } w \geq \frac{V}{2c}; \end{cases}$$

where $\mu_0 \in (\Lambda + c/V, 1/\omega)$ is the unique solution to $\lambda'(\mu_0) - \omega\mu_0 = 0$, where $\lambda(\mu) = \Lambda/(1 - \frac{-\sqrt{4k\mu(c+\Lambda V-\mu V)+(-k\Lambda+k\mu+\mu V)^2}-k\Lambda+k\mu+\mu V}{2k\mu})$. Moreover, $\mu^{FB} \geq \mu^{FIFO}$ and the inequality is strict if and only if $\omega < \bar{\omega}$.

In the first best, as scaling capacity gets easier, the firm builds more capacity. When it is difficult to scale capacity (i.e., $\omega \geq V/(2c)$), the firm should not launch the waitlist at all since the cost of serving customers would not justify the benefit of doing so. When it is moderately difficult to scale capacity (i.e., $\omega \in [\bar{\omega}, V/(2c))$), the firm should launch the waitlist but not the referral program. When it is easy to scale capacity (i.e., $w < \bar{\omega}$), it is in the firm's best interest to launch a referral program in conjunction with the waitlist. In this case, the firm maintains more capacity in the first best than it would with a non-referral FIFO queue. Intuitively, as referrals can increase the throughput for a given capacity, investing in capacity has more "bang for the buck" than doing so in a non-referral queue. Consequently, referrals prompt the firm to serve customers faster. Let $\underline{\Lambda} \triangleq \frac{ck}{(V-k)V}$ and $\bar{\Lambda} \triangleq \frac{ck}{(V-k)^2}$. Proposition 9 below characterizes the optimal structure of the referral program and the optimal capacity in the second best.

PROPOSITION 9. *If $\omega \in [\bar{\omega}, V/(2c))$, the first best is achieved by non-referral FIFO. If $\omega < \bar{\omega}$:*

(i) If $\Lambda \geq \bar{\Lambda}$ and $k < V$, the first best is achieved by partial priority;

(ii) If $\Lambda \in (\underline{\Lambda}, \bar{\Lambda})$ and $k < V$, then there exists a unique $\underline{\omega}$ such that the first best is achieved by partial priority if and only if $\omega \in (\underline{\omega}, \bar{\omega})$; the first best is achieved by full priority if $\omega = \underline{\omega}$; if $\omega < \underline{\omega}$, the first-best cannot be achieved and strategic delay is optimal in the second best;

(iii) If $\Lambda \leq \underline{\Lambda}$ and $k < V$, the first best cannot be achieved and there exists a unique $\hat{\omega}$ such that strategic delay is optimal if $\omega < \hat{\omega}$ and non-referral FIFO is optimal if $\omega \in [\hat{\omega}, \bar{\omega})$;

(iv) If $k \geq V$, the first best cannot be achieved and non-referral FIFO is optimal.

When the second best achieves the first best, the optimal second-best capacity, μ^{SB}, must statisfy $\mu^{SB} = \mu^{FB}$. When the second best does not achieve the first best, if the optimal structure in the second best is non-referral FIFO, then $\mu^{SB} = \mu^{FIFO}$; if it is strategic delay, then $\mu^{SB} = \frac{\Lambda(V-k)}{\omega[\Lambda(V-k)+c]} \in (\mu^{FIFO}, \mu^{FB})$. In general, $\mu^{FIFO} \leq \mu^{SB} \leq \mu^{FB}$. Further, whenever the second best generates referrals, the average delay in the second best is less than that in a non-referral FIFO queue.

Proposition 9 demonstrates the robustness of our insights from the second-best analysis when capacity is endogenously determined. If the first-best is achieved in the second best, the optimal structure of the mechanism is either partial/full priority or non-referral FIFO. If the first-best cannot be achieved (which generally occurs when it is easy for the firm to scale capacity but the market size is small), then either strategic delay or non-referral FIFO is optimal. In general, the second-best capacity is larger than the FIFO capacity, but smaller than first-best capacity. Similar to the argument for the first best, referrals accelerate customer acquisition also in the second best, and therefore the firm is willing to maintain a larger capacity than it would in a non-referral queue. However, the second best is not as efficient in generating referrals as the first best, and therefore the firm can be reluctant to build as much capacity as it would otherwise.

Interestingly, because the firm maintains more capacity when it launches a referral priority program, customers will experience less delay on average even though the referral program brings in more customers and potentially inserts strategic delay. Without the referral program, customers' expected delay cost is equal to their valuation of the service; with the referral program, however, their expected delay cost is reduced, equal to their valuation of the service less the expected referral cost. While the referral priority program leverages queueing delays as referral incentives, it surprisingly shortens how much customers wait once the firm optimally adjusts its speed. This result may give some reassurances to skeptics of the referral priority program.

6.4. Discriminating Between Base and Referred Customers

In our base model, the same expected-delay information $\boldsymbol{W}$ is provided to all customers impartially, regardless of their source (base or referred). As a result, base and referred customers adopt the

same joining and referral strategies. In practice, it may be cumbersome to quote different expected waiting time for different customers based on their source, and subsequently design scheduling policies to achieve those quoted expected delays. However, if the firm can employ a referral program that discriminates between base and referred customers, then there is a potential to further improve the throughput. We consider such an extension in this subsection.

Here is the description of the extended model. For each base customer who arrives, the firm announces expected delays (W_1^B, W_0^B), where W_1^B is the expected delay if that base customer makes a successful referral, and W_0^B is her expected delay if she does not. Likewise, for each referred customer who arrives, the firm announces expected delays (W_1^R, W_0^R), where W_1^R is the expected delay if that referred customer makes a successful referral, and W_0^R is her expected delay if she does not. Since base and referred customers are provided with potentially different expected-delay information, their joining and referral strategies can also differ. Let $(\alpha^B, \alpha^R, r^B, r^R)$ be a base customer's joining probability, a referred customer's joining probability, a base customer's referral effort and a referred customer's referral effort, respectively. The firm sets $(W_1^B, W_0^B, W_1^R, W_0^R)$ to maximize the system throughput[2], $\Lambda\alpha^B(1+r^B\alpha^R-r^B\alpha^R)/(1-r^B\alpha^R)$, subject to the IR, IC, OA and stability constraints. For the interest of space, we relegate the detailed mechanism design formulation to Appendix C. Note that in the base model, the expected-delay vector (W_1, W_0) is two-dimensional, and thus, we need $2^2-1=3$ (three) constraints to specify the operationally achievable region. By comparison, in the extended model, the expected-delay vector $(W_1^B, W_0^B, W_1^R, W_0^R)$ is four-dimensional, and thus, we need $2^4-1=15$ (fifteen) constraints to fully specify the operationally achievable region. Our base model is a special case of this extended model; setting $W_1^B = W_1^R$, $W_0^B = W_0^R$, $\alpha^B = \alpha^R$ and $r^B = r^R$ would reduce the extended model to the base one.

Table 4 The Optimal Mechanism That Discriminates Between Base and Referred Customers

Λ	0.85	0.7	0.01
Throughput	0.894	0.875	0.0841
α^B	1	1	1
α^R	1	1	1
r^B	0.0518	0.25	1
r^R	0	0	0.865
Structure	partial priority	partial priority	strategic delay

$V=10, c=1, k=2, \mu=1$.

Table 4 presents the outcome and structure of the optimal (discriminatory) mechanism for the model parameters used in Examples 1 through 3. Comparing the numerical results reported in

[2] The expression of the system throughput follows from Yang and Debo (2019).

Table 4 with those for the non-discriminatory case in Tables 1 through 3, we make the following observations. First, the structure of the optimal mechanism is preserved in these instances regardless of whether the firm discriminates. Specifically, when $\Lambda = 0.85$ or $\Lambda = 0.7$, the optimal structure is consistently partial priority, whereas the optimal structure is consistently strategic delay when $\Lambda = 0.01$. Second, similar to the non-discriminatory case, the optimal mechanism does not let the referral incentive crowd out the joining incentive (i.e., $\alpha^B = \alpha^R = 1$ for all three instances). Third, practicing discrimination further improves the firm's throughput across all three instances, although the improvement is marginal for the case of $\Lambda = 0.85$. Fourth, when $\Lambda = 0.85$ or $\Lambda = 0.7$ (corresponding to an intermediate base market size), the optimal mechanism only has base customers make referrals, but not the referred customers (i.e., $r^B > 0$ and $r^R = 0$). Finally, when $\Lambda = 0.01$ (corresponding to a small base market size), both base and referred customers refer in the optimal mechanism, but base customers exert more referral effort than referred ones (i.e., $r_B > r_R > 0$). Overall, these observations demonstrate the robustness of our insights and also suggest how a discriminatory policy, to the extent that it is feasible and easy to implement, can further accelerate customer acquisition.

7. Concluding Remarks

Using priorities on a waitlist as referral incentives, the referral priority program has emerged as an innovative business tool for customer acquisition. Despite its growing presence in the technology sector, design guidance is lacking as to how to configure the optimal referral mechanism. This paper proposes a unifying framework that rationalizes various referral schemes observed in practice, including non-referral FIFO, partial priority, full priority and strategic delay. Our results lend theoretical underpinnings to these practices and provide prescriptive insights that can advise waitlist managers on how to build the optimal referral priority program. While we have identified the ideal conditions for each type of referral programs to work, we caution that it may be precipitous to assume existing firms in practice have "figured it out" and are operating under optimality. If anything, our conversations with practitioners usually suggest otherwise: firms tend to have little clue as to what mechanism would best fit them, and they either indiscriminately experiment with different schemes, or rely on the default settings of a third-party waitlist tool (such as Waitlisted.co). The former can be unfruitful, whereas the latter can even be detrimental because as our results indicate, one size usually does not fit all; a referral program must be geared toward the specific business environment in which the firm is situated. In that regard, our paper would be instrumental in guiding firms' decisions.

Next, we discuss some modeling choices of the paper and future research directions. Our model makes the simplifying assumption that referrals convert instantly, whereas in reality, there could be

a time lag between the arrival of the referrer and that of the referred. Such a time lag would make the arrival process less bursty. Consequently, customers may have a stronger joining incentive (because of shorter delay induced by less burstiness) but a weaker referral incentive, because (1) shorter delay diminishes the value of gaining priority; and (2) referring customers may be served before her referred friend converts. As a result, we conjecture that the firm would be less concerned about the crowd-out effect of referrals, and would launch a referral program (as opposed to implementing FIFO) under a broader range of the base market size. However, motivating referrals would become harder, and therefore, the firm would increasingly gravitate toward the use of strategic delay.

While we assume for simplicity customers are taken off the waitlist one at a time, firms in practice often use batch processing, which would increase the burstiness of the service process (see, e.g., §4.6, Kleinrock 1975). Since burstiness lengthens delay and deters customers from joining, we conjecture that the first-best throughput found in our model would be an upper bound on the first-best throughput in a batch-service model with the same amount of capacity. On the other hand, as explained earlier, burstiness may strengthen referral incentives, and the optimal second-best mechanism may increasingly shift toward the use of partial priority and away from strategic delay. The above reasoning suggests that incorporating batch processing may partially offset the effect of incorporating a referral time lag.

In terms of whom to serve first, firms might be tempted to cherry-pick customers and release their services (or beta versions) first to those with high valuation or big influence (i.e., influencers). The referral priority program may help identify such customers as those who are more anxious about using the service or more influential in their social network may be more likely to refer friends and jump ahead in the queue. However, if cherry-picking degenerates into arbitrary service rules that no longer honor the expected delays posted as part of the referral program, then it can cause customer confusion and loss of goodwill.

As in most queueing-game papers, our model does not consider reneging. However, it is possible that a competing service is launched in the middle of a customer's wait and causes her to lose interest in the original service. One distinct feature of this setting is that "reneging" customers will not be removed from the waitlist even if they have technically abandoned the queue. The firm is generally uninformed of how many customers on the waitlist are still interested and how many are "phantoms." This raises an interesting question of how to retain customers while they are still on the waitlist, which nicely complements the current paper's focus on customer acquisition.

Our work focuses on the setting where customers are informed of the expected delay posted by the firm, but not the real-time queue length. As a result, our mechanism design problem restricts attention to static scheduling policies (typical in the queueing-game literature). In practice, customers' information structure varies from waitlist to waitlist. Some waitlists provide customers with

a wait-time estimate and some do not. Still, other waitlists further show the real-time queue length and inform each customer of their own positions. If the real-time queue length is disclosed, then one can consider finding the optimal dynamic, state-dependent scheduling policies for the referral program, which can be a daunting challenge. While the optimal dynamic mechanism in such an observable-queue is beyond the scope of the current study, the high-level trade-offs identified in this paper (such as the one between joining and referral incentives) still seem relevant and some of the structural properties we find about the static optimal referral program may still retain. Given the recent interest in information provision in queues (e.g., Allon et al. 2011, Hu et al. 2018, Lingenbrink and Iyer 2019, Wang and Hu 2019), an even more ambitious question for future research is how to jointly design dynamic information provision and scheduling policies. Note that in the era of social media, the amount of information a customer has may psychologically affect whether she signs up on a waitlist as many customers sign up presumably due to the fear of missing out, otherwise known as FOMO (McGinnis 2004). A long queue might be a social proof that prompts FOMO customers to sign up.

Finally, consistent with the literature, we assume that the firm has perfect knowledge of the demand (e.g., the base market size), whereas in practice (especially in our context), the firm may face demand uncertainty and only know the demand up to a certain precision, which would expose the firm to the risk of misspecification if it sets its referral program according to a demand scenario different from the true one. In light of such uncertainty, parameter-free mechanisms such as FIFO and full priority, may be viewed as more robust, and thus might be more practically appealing (as we pointed out earlier). Further adding to the complexity of the problem is that the business environment in practice may not always be stationary (as assumed in our model) but constantly evolving. For example, the base market size may grow as the firm's product or service trends upward. As a response, the firm may want to increase its capacity to keep up with the demand and adjust its referral program accordingly. In light of our current results, we conjecture that as the base market size grows, it may be in the firm's best interest to initially use strategic delay, later switch to partial priority, and eventually turn off the referral program. On a related note, our model implicitly assumes that the pool of potential customers to be referred is infinite. If the pool is finite, then an early user may easily identify a friend to refer while a latecomer may find it progressively more difficult to do so. One way to capture this effect is to assume the referral cost increases over time. Our current results suggest that as the referral cost increases, the firm should move from partial priority, to strategic delay and eventually to non-referral FIFO. A formal inquiry of referral priority programs in a non-stationary environment is left for future research.

Acknowledgments

The author thanks the department editor Morris Cohen, an anonymous associate editor, and three anonymous reviewers for their constructive feedback that has significantly improved the paper. The author is grateful to Laurens Debo, Justin McNally and Miles Wellesley for helpful discussion.

References

Afèche P (2013) Incentive-compatible revenue management in queueing systems: Optimal strategic delay. *Manufacturing & Service Operations Management* 15(3):423–443.

Afèche P, Mendelson H (2004) Pricing and priority auctions in queueing systems with a generalized delay cost structure. *Management Science* 50(7):869–882.

Afèche P, Pavlin M (2016) Optimal price-lead time menus for queues with customer choice: Priorities, pooling and strategic delay. *Management Science* 62(8):2412–2436.

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

Athreya KB, Ney PE (1972) *Branching Processes* (Berlin: Springer-Verlag).

Biyalogorsky E, Gerstner E, Libai B (2001) Customer referral management: Optimal reward programs. *Marketing Science* 20(1):82–95.

Business Insider (2013) There is a 260,000-person wait list for a new email app. Business Insider (February 7), URL `http://www.businessinsider.com/there-is-a-260000-person-wait-list-for-an-app-that-promises-to-fix-your-inbox-2013-2`.

Chen XH, Dempster AP, Liu JS (1994) Weighted finite population sampling to maximize entropy. *Biometrika* 81(3):457–469.

Coffman EJ, Mitrani I (1980) A characterization of waiting time performance realizable by single-server queues. *Operations Research* 28(3):810–821.

Cui S, Wang Z, Yang L (2019) The economics of line-sitting. *Management Science* URL `http://dx.doi.org/10.1287/mnsc.2018.3212`, article in advance.

Edelson N, Hildebrand D (1975) Congestion tolls for Poisson queuing processes. *Econometrica* 43(1):81–92.

Financial Times (2018) The rise of AI and remote assistants. Financial Times (January 28), URL `https://www.ft.com/content/6591a6fc-f7cf-11e7-a4c9-bbdefa4f210b`.

Fortune (2015) How Robinhood, an investing app, is luring stock-market newbies. Fortune (March 12), URL `http://fortune.com/2015/03/12/robinhood-investing-app/`.

Gavirneni S, Kulkarni VG (2016) Self-selecting priority queues with Burr distributed waiting costs. *Production and Operations Management* 25(6):979–992.

Glazer A, Hassin R (1986) Stable priority purchasing in queues. *Operations Research Letter* 4(6):285–288.

Gurvich I, Lariviere M, Ozkan C (2019) Coverage, coarseness and classification: Determinants of social efficiency in priority queues. *Management Science* 65(3):1061–1075.

Harremoes P (2001) Binomial and Poisson distributions as maximum entropy distributions. *IEEE Transactions on Information Theory* 47(5):2039–2041.

Hassin R (1995) Decentralized regulation of a queue. *Management Science* 41(1):163–173.

Hassin R (2016) *Rational Queueing* (Boca Raton: CRC Press, Taylor and Francis Group).

Hassin R, Haviv M (2003) *To Queue Or Not to Queue: Equilibrium Behavior in Queueing Systems* (Boston: Kluwer Academic Publishers).

Holmström B (1979) Moral hazard and observability. *Bell Journal of Economics* 10(1):74–91.

Hu M, Li Y, Wang J (2018) Efficient ignorance: Information heterogeneity in a queue. *Management Science* 64(6):2650–2671.

Jing X, Xie J (2011) Group buying: A new mechanism for selling through social interactions. *Management Science* 57(8):1354–1372.

Katta A, Sethuraman J (2005) Pricing strategies and service differentiation in queues: a profit maximization perspective. Working paper, Columbia University.

Kleinrock L (1967) Optimum bribing for queue position. *Operations Research* 15(2):304–318.

Kleinrock L (1975) *Queueing Systems. Volume 1: Theory* (New York: Wiley-Interscience).

Kornish LJ, Li Q (2010) Optimal referral bonuses with asymmetric information: Firm-offered and interpersonal incentives. *Marketing Science* 29(1):108–121.

Lingenbrink D, Iyer K (2019) Optimal signaling mechanisms in unobservable queues. *Operations Research* 67(5):1397–1416.

Lobel I, Sadler E, Varshney LR (2017) Customer referral incentives and social media. *Management Science* 63(10):3514–3529.

Lui FT (1985) An equilibrium queueing model of bribery. *Journal of Political Economy* 93(4):760–781.

Maglaras C, Yao J, Zeevi A (2018) Optimal price and delay differentiation in large-scale queueing systems. *Management Science* 64(5):2427–2444.

McGinnis PJ (2004) Social theory at HBS: McGinnis' Two FOs. The Harbus (May 10), URL `http://harbus.org/2004/social-theory-at-hbs-2749/`.

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

Myerson RB (1981) Optimal auction design. *Mathematics of Operations Research* 6(1):58–73.

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

Ross SA (1973) The economic theory of agency: The principal's problem. *American Economic Review* 62(2):134–139.

Su X, Zenios SA (2006) Recipient choice can address the efficiency-equity trade-off in kidney transplantation: A mechanism design model. *Management Science* 52(11):1647–1660.

TechCrunch (2011) How DropBox started as a minimal viable product. TechCrunch (October 19), URL `https://techcrunch.com/2011/10/19/dropbox-minimal-viable-product/`.

The Verge (2013) Expect delays: why today's top apps are putting you on a wait list. The Verge (July 30), URL `http://www.theverge.com/2013/7/30/4567794/mailbox-loom-cloud-app-wait-lists`.

Wang J, Cui S, Wang Z (2019) Equilibrium strategies in $M/M/1$ priority queues with balking. *Production and Operations Management* 28(1):43–62.

Wang J, Hu M (2019) Efficient inaccuracy: User-generated information sharing in a queue. *Management Science* Forthcoming.

Wang YH (1993) On the number of successes in independent trials. *Statistica Sinica* 3(2):295–312.

Xiao P, Tang CS, Wirtz J (2011) Optimizing referral reward programs under impression management considerations. *European Journal of Operational Research* 215(3):730–739.

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

Yang L, Debo L, Gupta V (2017) Trading time in a congested environment. *Management Science* 63(7):2377–2395.

Yang L, Wang Z, Cui S (2019) A model of queue-scalping. Working paper, Johns Hopkins University.

Online Appendix to "Invite Your Friend and You'll Move up in Line: Optimal Design of Referral Priority Programs"

Luyi Yang

Appendix A: Proofs

Proof of Proposition 1 The first-best problem relaxes IC constraint (5) in Problem 1 and effectively reduces to: $\max_{\alpha,r} \frac{\Lambda\alpha}{1-r\alpha}$, s.t. $V - rk - \frac{c}{\mu(1-r\alpha)-\Lambda\alpha} \geq 0$. Let $\gamma \geq 0$, be the Lagrangian multiplier for the constraint. Thus, the Lagrangian becomes $\mathcal{L} = \frac{\Lambda\alpha}{1-r\alpha} + \gamma\left(V - rk - \frac{c}{\mu(1-r\alpha)-\Lambda\alpha}\right)$. From the KKT conditions $\frac{\partial\mathcal{L}}{\partial r} = \frac{\Lambda\alpha^2}{(1-r\alpha)^2} + \gamma\left(-k - c\frac{\mu\alpha}{[\mu(1-r\alpha)-\Lambda\alpha]^2}\right) = 0$ and $\frac{\partial\mathcal{L}}{\partial\alpha} = \frac{\Lambda}{(1-r\alpha)^2} + \gamma\left(-c\frac{\mu r+\Lambda}{[\mu(1-r\alpha)-\Lambda\alpha]^2}\right) = 0$, we have $\gamma > 0$. By complementary slackness, $V - rk - \frac{c}{\mu(1-r\alpha)-\Lambda\alpha} = 0$. This implies that $r\alpha W_1 + (1-r\alpha)W_0 = \frac{1}{\mu(1-r\alpha)-\Lambda\alpha}$. Hence, the optimal scheduling policy must always be work-conserving in the first-best problem. From $U \equiv V - rk - \frac{c}{\mu(1-r\alpha)-\Lambda\alpha} = 0$, $\alpha(r) \triangleq \frac{\mu(V-rk)-c}{(V-rk)(\Lambda+\mu r)} \leq 1$. Also, $\partial U/\partial\alpha < 0$ and $\partial U/\partial r < 0$. Thus, $d\alpha/dr = -(\partial U/\partial r)/(\partial U/\partial\alpha) < 0$; α is decreasing in r. The objective function can be expressed as a function of r: $\frac{\Lambda\alpha}{1-r\alpha} = \frac{\Lambda\frac{\mu(V-rk)-c}{(V-rk)(\Lambda+\mu r)}}{1-r\frac{\mu(V-rk)-c}{(V-rk)(\Lambda+\mu r)}} = \Lambda\frac{\mu(V-rk)-c}{\Lambda(V-rk)+cr}$. We shall show this function is decreasing in r by showing its reciprocal is increasing in r. It suffices to show $\frac{\Lambda(V-rk)+cr}{\mu(V-rk)-c} = \frac{\Lambda}{\mu} + \frac{cr+\Lambda c/\mu}{\mu(V-rk)-c}$ is increasing in r, which is obvious. Therefore, the objective function is decreasing in r. Note that $\alpha(0) = \frac{\mu V-c}{V\Lambda}$. Thus, $\alpha(0) \leq 1$ if and only if $\Lambda \geq \mu - c/V$. Thus, when $\Lambda \geq \mu - c/V$, $r = 0$ and $\alpha = \frac{\mu V-c}{V\Lambda}$. When $\Lambda < \mu - c/V$, $\alpha = 1$ and $r \in (0,1)$ solves $\alpha(r) = 1$, or $\frac{\mu(V-rk)-c}{(V-rk)(\Lambda+\mu r)} = 1$. Collecting terms gives $\mu - \frac{c}{V-rk} - \mu r = \Lambda$. It is easy to see $\partial r/\partial\Lambda < 0$. Solving for r gives $r = \frac{-\sqrt{4k\mu(c+\Lambda V-\mu V)+(-k\Lambda+k\mu+\mu V)^2}-k\Lambda+k\mu+\mu V}{2k\mu}$. □

Proof of Proposition 3 Any mechanism that generate referrals must have $k \leq c\alpha(W_0 - W_1)$, or equivalently, $W_0 \geq k/(c\alpha) + W_1$. Moreover, the OA constraint requires $W_1 \geq 1/(\mu(1-r\alpha) - \Lambda r\alpha^2)$ and the IR constraint requires $V - rk - c[r\alpha W_1 + (1-r\alpha)W_0] \geq 0$. Thus, by substitution, any mechanism that generates referrals must have

$$V - \frac{k}{\alpha} - \frac{c}{\mu(1-r\alpha) - \Lambda r\alpha^2} \geq 0. \tag{A.1}$$

Since $\frac{c}{\mu(1-r\alpha)-\Lambda r\alpha^2} > \frac{c}{\mu}$ and $\frac{k}{\alpha} \geq k$, thus to generate referrals, we must have $V - k - \frac{c}{\mu} > 0$. Therefore, if $V - c/\mu \leq k$, no referrals can be generated. □

Proof of Proposition 4 In the first-best problem, when $\Lambda < \mu - c/V$, $\alpha^{FB} = 1$ and r^{FB} solves

$$V - r^{FB}k - \frac{c}{\mu(1-r^{FB}) - \Lambda} = 0. \tag{A.2}$$

It is achievable by the second best if and only if W_1 and W_0 satisfy $r^{FB}W_1 + (1-r^{FB})W_0 = \frac{1}{\mu(1-r^{FB})-\Lambda}$, $k = c[W_0 - W_1]$, $W_1 \geq \frac{1}{\mu(1-r^{FB})-\Lambda r^{FB}}$. The first equation is due to work-conservation of the first-best solution. The second equation is due to $r^{FB} \in (0,1)$, i.e., customers should be indifferent to referrals. The inequality is from the OA constraint. Solving the first two equations gives $W_1 = \frac{1}{\mu(1-r^{FB})-\Lambda} - \frac{k(1-r^{FB})}{c}$, $\quad W_0 = \frac{1}{\mu(1-r^{FB})-\Lambda} + \frac{kr^{FB}}{c}$; Thus, we need to show

$$\frac{1}{\mu(1-r^{FB}) - \Lambda} - \frac{k(1-r^{FB})}{c} \geq \frac{1}{\mu(1-r^{FB}) - \Lambda r^{FB}} \tag{A.3}$$

if and only if $\Lambda \geq \Lambda^*$ (subject to $\Lambda \geq \mu - c/V$), where $\Lambda^* \in (0, \mu - c/V)$ is uniquely determined. Combining (A.2) and (A.3), it follows that (A.3) holds if and only if

$$V - k - \frac{c}{\mu(1-r^{FB}) - \Lambda r^{FB}} \geq 0. \tag{A.4}$$

From (A.2), $\Lambda = \mu(1-r^{FB}) - \frac{c}{V - r^{FB}k}$. Note that Λ is decreasing in r^{FB}. Plugging it into (A.4) gives $V - k - \frac{c}{\mu(1-r^{FB}) - \Lambda r^{FB}} = V - k - \frac{c}{\mu(1-r^{FB}) - \mu(1-r^{FB})r^{FB} + r^{FB}\frac{c}{V - r^{FB}k}} = V - k - \frac{c}{\mu(1-r^{FB})^2 + \frac{r^{FB}c}{V - r^{FB}k}} \triangleq f(r^{FB})$. Thus, proving the proposition is equivalent to showing that when $V - c/\mu > k$, there exists a unique $r^* \in (0,1)$ such that $f(r^*) = 0$. Moreover, $f(r) > 0$ if $r \in (0, r^*)$ and $f(r) < 0$ if $r \in (r^*, 1)$. By inspection, $f(1) = 0$; $f(0) = V - k - c/\mu > 0$. $f(r) = V - k - \frac{c(V-rk)}{\mu(1-r)^2(V-rk) + rc}$. Since $\mu(1-r)^2(V-rk) + rc > 0$ for $r \in (0,1)$, it suffices to show $g(r) \triangleq f(r)(\mu(1-r)^2(V-rk) + rc)$ has this property. Plugging $f(r)$ into $g(r)$ gives $g(r) = (\mu(1-r)^2(V-rk) + rc)(V-k) - c(V-rk)$. Note that $g(r)$ is a third-degree polynomial function, which has at most three roots. We already know one root $r = 1$. The coefficient for r^3 is $-k\mu(V-k) < 0$. Therefore, $g(r) < 0$ for sufficiently large r. The derivative of $g(r)$ evaluated at $r = 1$ is $g'(1) = cV > 0$. This implies there exists $\varepsilon > 0$ such that $g(1+\varepsilon) > 0$ and $g(1-\varepsilon) < 0$. Since $g(0) > 0$ and $g(1-\varepsilon) < 0$, there exists $r^* \in (0,1)$ such that $g(r^*) = 0$ by the intermediate value theorem. Likewise, since $g(1+\varepsilon) > 0$ and $g(r) < 0$ for sufficiently large r, there also exists an $r^{*\prime} > 1$ such that $g(r^{*\prime}) = 0$. Since $g(r)$ has at most three roots and one root is 1, we conclude that r^* and $r^{*\prime}$ must be unique. Moreover, $g(r) > 0$ if $r \in (0, r^*)$ and $g(r) < 0$ if $r \in (r^*, 1)$. Therefore, f also has this property. We also need to show $\Lambda^* = \mu(1-r^*) - \frac{c}{V - r^*k} > 0$. Since Λ is decreasing in r, It suffices to show $f(r_0) < 0$, where r_0 solves $\mu(1-r_0) - \frac{c}{V - r_0 k} = 0$. By definition, $f(r_0) = V - k - \frac{c}{\mu(1-r_0)} = V - k - \frac{c}{\frac{c}{V - r_0 k}} = -(1-r_0)k < 0$. Therefore, $\Lambda^* > 0$. Finally, solving $g(r^*) = 0$ for r^* gives $r^* = \frac{\sqrt{(\mu V^2 - k^2\mu)^2 - 4(k^2\mu - k\mu V)(cV + k\mu V - \mu V^2)} + k^2\mu - \mu V^2}{2(k^2\mu - k\mu V)}$ and $\Lambda^* = \left[\mu(1-r^*) - \frac{c}{V - r^*k}\right]$.

Next, we identify the optimal second-best mechanism for $\Lambda < \Lambda^*$ and $k < V - c/\mu$. Either $r = 0$ (FIFO) or $r > 0$ (referrals generated). Now, we compare these two candidate mechanisms.

Case 1: $r = 0$. Since $\Lambda < \Lambda^* < \mu - c/V$, $V - c/(\mu - \Lambda) > 0$; thus, $\alpha = 1$, and the system throughput is Λ.

Case 2: $r > 0$. From the proof of Proposition 3, any mechanism that generates referrals must satisfy (A.1). Thus, the optimal value to the following relaxation problem is an upper bound of the system throughput with $r > 0$: $\max_{\alpha, r} \ \frac{\Lambda\alpha}{1 - r\alpha}$ s.t. $V - \frac{k}{\alpha} - \frac{c}{\mu(1-r\alpha) - \Lambda r\alpha^2} \geq 0$. Note that at this point, we have not shown that the solution of this problem is implementable. We will show this later. In this relaxation problem, the constraint must be binding because otherwise one can always increase r to increase the objective function while still satisfying the constraint. Therefore, $V - k/\alpha - \frac{c}{\mu(1-r\alpha) - \Lambda r\alpha^2} = 0$, which gives $r = \frac{-\alpha c + \alpha\mu V - k\mu}{\alpha(\alpha V - k)(\alpha\Lambda + \mu)}$. Plugging it into the objective function gives $\Lambda\frac{\alpha}{1-r\alpha} = \Lambda\frac{(\alpha V - k)(\alpha\Lambda + \mu)}{\alpha\Lambda V + c - k\Lambda} = \Lambda\frac{\alpha\Lambda + \mu}{\Lambda + \frac{c}{\alpha V - k}}$ From $r = \frac{-\alpha c + \alpha\mu V - k\mu}{\alpha(\alpha V - k)(\alpha\Lambda + \mu)}$, $\alpha V - k = \frac{c\alpha}{\mu(1-r\alpha) - \Lambda r\alpha^2} > 0$. Therefore, $\Lambda\alpha/(1 - r\alpha) = \Lambda\frac{\alpha\Lambda + \mu}{\Lambda + \frac{c}{\alpha V - k}}$ is increasing in α. Hence, the optimal $\alpha^{\mathrm{SB}} = 1$. Plugging $\alpha = 1$ into $r = \frac{-\alpha c + \alpha\mu V - k\mu}{\alpha(\alpha V - k)(\alpha\Lambda + \mu)}$ gives $r^{\mathrm{SB}} = \frac{\mu(V-k) - c}{(V-k)(\mu + \Lambda)}$. Note that the throughput in this relaxation problem is obviously higher than the FIFO throughput Λ in Case 1. Since $V - k/\alpha - \frac{c}{\mu(1-r\alpha) - \Lambda r\alpha^2} = 0$, it follows from the logic of the proof of Proposition 3 that $W_1 = \frac{1}{\mu(1-r) - \Lambda r}, \quad W_0 = W_1 + k/c$. It remains to be shown that the solution to this relaxation problem is implementable. We check if all the constraints in the mechanism design problem are satisfied. The IC constraint is satisfied by construction since $c\alpha(W_0 - W_1) = k$. The IR constraint is satisfied because

$$V - rk - c[r\alpha W_1 - (1 - r\alpha)W_0] = V - k - \frac{c}{\mu(1-r) - \Lambda r} = 0. \tag{A.5}$$

The OA constraint $W_1 \geq 1/(\mu(1-r\alpha)-\Lambda r\alpha^2)$ is obviously satisfied. We check the other OA constraint: $r\alpha W_1 + (1-r\alpha)W_0 \geq 1/(\mu(1-r\alpha)-\Lambda\alpha)$. From (A.5), it is equivalent to checking

$$V - rk - \frac{c}{\mu(1-r)-\Lambda} \geq 0. \tag{A.6}$$

We shall show that this is true and the inequality is strict if $\Lambda < \Lambda^*$, which would prove the optimality of strategic delay. First note that by definition of Λ^*, at Λ^*, $V - r^*k - \frac{c}{\mu(1-r^*\alpha)-\Lambda^*\alpha} = 0$, where r^* solves $V - k - \frac{c}{\mu(1-r^*)-\Lambda r^*} = 0$. From (A.5), $\Lambda = \mu(1-r)/r - \frac{c}{(V-k)r}$. Moreover, decreasing Λ increases r, which implies that for $\Lambda < \Lambda^*$, $r > r^*$. Plugging this into the LHS of (A.6) gives $V - rk - \frac{c}{\mu(1-r)-\Lambda} = V - rk - \frac{c}{\mu(1-r)-\mu(1-r)/r+\frac{c}{(V-k)r}} = V - rk - \frac{c(V-k)r}{c-(V-k)\mu(1-r)^2} \triangleq \phi(r)$. To prove (A.6), it suffices to show $\phi(r) > 0$ for $r \in (r^*, 1)$. Note that $c-(V-k)\mu(1-r)^2 > 0$ since $\frac{c(V-k)r}{c-(V-k)\mu(1-r)^2} = \frac{c}{\mu(1-r)-\Lambda} > 0$ and $c(V-k)r > 0$. Therefore, it is equivalent to showing $(V-rk)(c-(V-k)\mu(1-r)^2) - c(V-k)r > 0$ for $r \in (r^*, 1)$. Recognize that the LHS is equal to $-g(r)$, where $g(r)$ is defined in the proof of Proposition 4, in which we show $g(r) < 0$ for $r > r \in (r^*, 1)$. Therefore, $-g(r) > 0$ for $r > r \in (r^*, 1)$. This shows that the solution to the relaxation problem is implementable, and since it achieves a higher throughput than FIFO in Case 1. This (strategic-delay) solution is optimal. $\square$

Proof of Proposition 5 The pair (Λ^*, r^*) solves the following joint equations:

$$V - r^*k - \frac{c}{\mu(1-r^*)-\Lambda^*} = 0, \quad V - k - \frac{c}{\mu(1-r^*)-\Lambda r^*} = 0. \tag{A.7}$$

Expressing Λ^* as a function of Λ from the first equation of (A.7) and plugging it into the second equation of (A.7) gives $g(r^*) = [\mu(1-r^*)^2(V-r^*k)+r^*c](V-k) - cV + cr^*k = 0$. Note that $g(\cdot)$ is the same g function defined in the proof of Proposition 4. Simplifying $g(\cdot)$ yields $g(r) = (1-r)[\mu(1-r)(V-rk)(V-k) - cV] = 0$. Since $r^* \in (0,1)$, $h(r^*) = \mu(1-r^*)(V-r^*k)(V-k) - cV = 0$. By inspection, $\frac{\partial h(r)}{\partial r} < 0$, $\frac{\partial h}{\partial k} < 0$. Therefore, $\frac{\partial r^*}{\partial k} = -\frac{\partial h/\partial k}{\partial h/\partial r} < 0$. Next, we shall show $\partial\Lambda^*/\partial r^* < 0$, which would prove $\partial\Lambda^*/\partial k > 0$. From the second equation of (A.7), $k = V - c/(\mu(1-r^*)-\Lambda r^*)$. Plugging it into the first equation of (A.7) gives $V - r^*\left(V - \frac{c}{\mu(1-r^*)-\Lambda r^*)}\right) - \frac{c}{\mu(1-r^*)-\Lambda^*} = 0$. Collecting terms gives $V(1-r^*) - \frac{c}{\mu(1-r^*)-\Lambda^*} + \frac{cr^*}{\mu(1-r^*)-\Lambda^* r^*} = 0$. $V(1-r^*) - \frac{c\mu(1-r^*)^2}{(\mu(1-r^*)-\Lambda^*)(\mu(1-r^*)-\Lambda^* r^*)} = 0$. $\tau(\Lambda^*, r^*) = V(1-r^*) - \frac{c\mu}{(\mu-\Lambda^*/(1-r^*))(\mu-\Lambda^* r^*/(1-r^*))} = 0$. By inspection, $\frac{\partial\tau}{\partial r^*} < 0$, $\frac{\partial\tau}{\partial\Lambda^*} < 0$. Therefore, $\frac{\partial\Lambda^*}{\partial r^*} = -\frac{\partial\tau/\partial r^*}{\partial\tau/\partial\Lambda^*} < 0$. Since $\frac{\partial\Lambda^*}{\partial r^*} < 0$ and $\frac{\partial r^*}{\partial k} < 0$, by the chain rule, $\frac{\partial\Lambda^*}{\partial k} = \frac{\partial\Lambda^*}{\partial r^*}\frac{\partial r^*}{\partial k} > 0$. $\square$

Proof of Proposition 6 The first-best problem can simplify to $\max_{\boldsymbol{r}\in[0,1]^D, \alpha\in[0,1]} \Pi(\boldsymbol{r},\alpha) = \frac{\Lambda\alpha}{1-m(\boldsymbol{r},\alpha)}$, s.t. $U(\boldsymbol{r},\alpha) = V - k\sum_{i=1}^D r_i - c\frac{\sigma^2(\boldsymbol{r},\alpha)+(2-m(\boldsymbol{r},\alpha))(1-m(\boldsymbol{r},\alpha))}{2(1-m(\boldsymbol{r},\alpha))[\mu(1-m(\boldsymbol{r},\alpha))-\Lambda\alpha]} \geq 0$, where $m(\boldsymbol{r},\alpha) = \sum_{i=1}^D r_i\alpha$, and $\sigma^2(\boldsymbol{r},\alpha) = \sum_{i=1}^D r_i\alpha(1-r_i\alpha)$. To prove Proposition 6, we need to prove that any feasible solution to the first-best problem $(\boldsymbol{r}^*, \alpha^*)$ with two or more positive r_i^*'s can be strictly dominated by a feasible solution with at most one positive $r_i > 0$. Let $m^* = \sum_{i=1}^D r_i^*\alpha^*$. Then, $U(\boldsymbol{r}^*, \alpha^*) = V - k\sum_{i=1}^D r_i^* - c\frac{\sum_{i=1}^D r_i^*\alpha^*(1-r_i^*\alpha^*)+(2-m^*)(1-m^*)}{2(1-m^*)[\mu(1-m^*)-\Lambda\alpha^*]} \geq 0$. We first prove the following lemma:

LEMMA A.1. *For any $K \in (0,1)$, the minimization problem* $\min_{x_1,\ldots,x_D\in[0,1], x_1\geq x_2\geq\ldots\geq x_D} \sum_{i=1}^D x_i(1-x_i)$ s.t. $\sum_{i=1}^D x_i = m$ *has the optimal value $m(1-m)$ with the unique optimal solution $x_1 = m$ and $x_2 = x_3 = \ldots = x_D = 0$.*

Proof of Lemma A.1 To see this, note $\sum_{i=1}^{D} x_i(1-x_i) = \sum_{i=1}^{D} x_i - \sum_{i=1}^{D} x_i^2 = m - \left(\sum_{i=1}^{D} x_i\right)^2 + 2\sum_{i,j:i\neq j} x_i x_j = m - m^2 + 2\sum_{i,j:i\neq j} x_i x_j \geq m(1-m)$. On the other hand, $x_1 = m$ and $x_2 = x_3 = \ldots = x_D = 0$ gives $\sum_{i=1}^{D} x_i(1-x_i) = m(1-m)$. Moreover, $2\sum_{i,j:i\neq j} x_i x_j > 0$ if there are two positive x_i's, which implies any solution with two positive x_i's cannot achieve the objective value $m(m-1)$. Hence, $x_1 = m$ and $x_2 = x_3 = \ldots = x_D = 0$ is the unique optimal solution and $m(1-m)$, the optimal value. □

Lemma A.1 implies that for any $(\boldsymbol{r}^*, \alpha^*)$ with two or more positive r_i^*'s, $\frac{\sum_{i=1}^{D} r_i^*\alpha^*(1-r_i^*\alpha^*)+(2-m^*)(1-m^*)}{2(1-m^*)[\mu(1-m^*)-\Lambda\alpha^*]} > \frac{m^*(1-m^*)+(2-m^*)(1-m^*)}{2(1-m^*)[\mu(1-m^*)-\Lambda\alpha^*]} = \frac{1}{\mu(1-m^*)-\Lambda\alpha^*}$. Moreover, since $U(\boldsymbol{r}^*, \alpha^*) \geq 0$, we have

$$V - k\frac{m^*}{\alpha^*} - \frac{c}{\mu(1-m^*)-\Lambda\alpha^*} > 0. \tag{A.8}$$

Now, we discuss two cases.

Case 1: $\alpha^*/(1-m^*) < 1$. Let $\alpha' = \alpha^*/(1-m^*)$. Thus, $U_1(\alpha') \equiv V - \frac{c}{\mu-\Lambda\alpha'} = V - \frac{c(1-m^*)}{\mu(1-m^*)-\Lambda\alpha^*} > V - \frac{c}{\mu(1-m^*)-\Lambda\alpha^*} > 0$, where the last inequality follows from (A.8). Since $U_1(\alpha')$ is decreasing in α', there must exist a unique solution $\alpha'' > \alpha'$ such that $U_1(\alpha'') = 0$. Let $\alpha''' = \min\{\alpha'', 1\}$. Thus, $\Lambda\alpha''' > \Lambda\alpha^*/(1-m^*)$. This implies that a non-referral solution $(\boldsymbol{r}, \alpha) = (\boldsymbol{0}, \alpha''')$ can strictly dominate $(\boldsymbol{r}^*, \alpha^*)$ with two or more positive r_i^*'s, and therefore, the latter is not an optimal solution to the first-best problem.

Case 2: $\alpha^*/(1-m^*) \geq 1$. Let $r_1' = 1 - (1-m^*)/\alpha^* \in [0,1)$. $U_2(r_1') = V - kr_1' - \frac{c}{\mu(1-r_1')-\Lambda} = V - k\left(1 - \frac{1}{\alpha^*} + \frac{m^*}{\alpha^*}\right) - \frac{c\alpha^*}{\mu(1-m^*)-\Lambda\alpha^*} > V - k\frac{m^*}{\alpha^*} - \frac{c}{\mu(1-m^*)-\Lambda\alpha^*} > 0$, where the last inequality follows from (A.8). Since $U_2(r_1')$ is decreasing in r_1', there must exist a unique solution $r_1'' \in (r_1', 1)$ such that $U_2(r_1'') = 0$. Note that $\Lambda/(1-r_1'') > \Lambda/(1-r_1') = \Lambda\alpha^*/(1-m^*)$. This implies that a new solution $(\boldsymbol{r}, \alpha)$ with $r_1 = r_1''$, $r_2 = r_3 = \ldots r_D = 0$ and $\alpha = 1$ can strictly dominate $(\boldsymbol{r}^*, \alpha^*)$ with two or more positive r_i^*'s, and therefore, the latter is not an optimal solution to the first-best problem. Combining the two cases shows that the first-best problem can have at most one positive r_i in the optimal solution. □

Proof of Proposition 7 Under the transfer mechanism proposed, we show that (α^{FB}, r^{FB}) satisfies both IR and IC and thus is an equilibrium outcome. If customers adopt strategy (α, r), then each customer' expected utility is $U(r,\alpha) = \alpha\{V - rk + r\alpha P - c[r\alpha W_1(r,\alpha) + (1-r\alpha)W_0(r,\alpha)] - r^{FB}P\}$, where $W_1(r,\alpha) = 1/[\mu(1-r\alpha) - \Lambda r\alpha^2]$ and $r\alpha W_1(r,\alpha) + (1-r\alpha)W_0(r,\alpha) = 1/(\mu(1-r\alpha) - \Lambda\alpha)$. Thus, $U(r^{FB}, \alpha^{FB} = 1) = V - r^{FB}k - \frac{c}{\mu(1-r^{FB})-\Lambda} = 0$. Hence, the IR constraint is satisfied by (α^{FB}, r^{FB}). To satisfy the IC constraint, we must have $k = c\alpha[W_0(r^{FB}, \alpha^{FB}) - W_1(r^{FB}, \alpha^{FB})] + P$, which is satisfied by construction. Thus, the proposed transfer mechanism can induce the first-best customer outcome. Also, note that it is budget-balanced and therefore, it achieves the first-best system throughput without any (long-run) financial cost/gain for the firm. Note that when the first-best is not achieved in the base model, we must have $V - k - c/[\mu(1-r^{FB}) - \Lambda r^{FB}] < 0$; since $V - r^{FB}k - \frac{c}{\mu(1-r^{FB})-\Lambda} = 0$, we must have $P = k - c\frac{\Lambda}{[\mu(1-r^{FB})-\Lambda][\mu(1-r^{FB})-\Lambda r^{FB}]} > 0$. □

Proof of Proposition 8 We first solve for the optimal FIFO capacity. Given capacity μ, the expected utility from joining a FIFO queue is $V - c/(\mu - \Lambda\alpha)$. Thus, the system throughput $\lambda^{FIFO}(\mu)$ is: $\lambda^{FIFO}(\mu) = 0$, if $\mu < c/V$; $\lambda^{FIFO}(\mu) = \mu - c/V$, if $c/V \leq \mu \leq \Lambda + c/V$; $\lambda^{FIFO}(\mu) = \Lambda$, if $\mu > \Lambda + c/V$. The optimal objective value is $\Pi^{FIFO} = [\max_\mu \lambda^{FIFO}(\mu) - \omega\mu^2/2]^+$. Since $V > c/\Lambda$, we have $\Pi^{FIFO} = \Lambda - \frac{1}{2}\omega(\Lambda + c/V)^2$, if $\omega < V/(\Lambda V + c)$; $\Pi^{FIFO} = -c/V + \frac{1}{2\omega}$, if $\omega \in [V/(\Lambda V + c), V/(2c))$; $\Pi^{FIFO} = 0$, if $\omega \geq V/(2c)$. The corresponding optimal

FIFO capacity μ^{FIFO} is $\mu^{FIFO}=\Lambda+c/V$, if $w<\frac{V}{\Lambda V+c}$; $\mu^{FIFO}=1/\omega$, if $w\in[\frac{V}{\Lambda V+c},\frac{V}{2c})$; $\mu^{FIFO}=0$, if $w\geq\frac{V}{2c}$. Next, we characterize the first best of the referral program. By Proposition 1, in the first best, the system throughput as a function of capacity μ, $\lambda^{FB}(\mu)$, is $\lambda^{FB}(\mu)=0$, if $\mu<c/V$; $\lambda^{FB}(\mu)=\mu-c/V$, if $c/V\leq\mu\leq\Lambda+c/V$; $\lambda^{FB}(\mu)=\Lambda/(1-r^{FB}(\mu))$, if $\mu>\Lambda+c/V$; where $r^{FB}(\mu)$ solves $V-kr^{FB}(\mu)-\frac{c}{\mu(1-r^{FB}(\mu))-\Lambda}=0$. Note that $r^{FB}(\mu)$ is increasing in μ. Thus, $\lambda^{FB}(\mu)=\Lambda/[1-r^{FB}(\mu)]$ is increasing in μ. Next, we show that $\lambda^{FB}(\mu)$ is concave in μ for $\mu>\Lambda+c/V$. Plugging $r^{FB}(\mu)=1-\Lambda/\lambda^{FB}$ into $V-kr^{FB}(\mu)-\frac{c}{\mu(1-r^{FB}(\mu))-\Lambda}=0$ gives $V-k+k\frac{\Lambda}{\lambda^{FB}}-\frac{c\lambda^{FB}}{\Lambda(\mu-\lambda^{FB})}=0$. Writing μ as a function of λ^{FB} gives $\mu=\frac{c\lambda^{FB}}{\Lambda[V-k+k\frac{\Lambda}{\lambda^{FB}}]}+\lambda^{FB}$. Note that $V-k+k\frac{\Lambda}{\lambda^{FB}}>0$, which is equivalent to $\lambda^{FB}V-\lambda^{FB}k+k\Lambda>0$.

$$\frac{d\mu}{d\lambda^{FB}}=1+\frac{c\lambda^{FB}(\lambda^{FB}V-k\lambda^{FB}+2k\Lambda)}{\Lambda(k(-\lambda^{FB}+\Lambda)+\lambda^{FB}V)^2}>1;\quad \frac{d^2\mu}{d(\lambda^{FB})^2}=\frac{2ck^2\Lambda}{(\lambda^{FB}V-k\lambda^{FB}+k\Lambda)^3}>0. \tag{A.9}$$

Therefore, μ is convex increasing in λ^{FB}. From Mršević (2008), the inverse of a convex increasing function is a concave increasing function, i.e., λ^{FB} is concave increasing in μ. Since the capacity cost is $\omega\mu^2/2$ per unit time, the first order condition of the objective function is $\Pi'(\mu)=d\lambda^{FB}(\mu)/d\mu-\omega\mu$. Since λ^{FB} is concave in μ for $\mu>\Lambda+c/V$, $\Pi'(\mu)$ is decreasing in μ for $\mu>\Lambda+c/V$. It follows from the equation $\mu=\frac{c\lambda^{FB}}{\Lambda[V-k+k\frac{\Lambda}{\lambda^{FB}}]}+\lambda^{FB}$ that when $\mu=\Lambda+c/V$, $\lambda^{FB}=\Lambda$. From (A.9), $\frac{d\mu}{d\lambda^{FB}}\big|_{\lambda^{FB}=\Lambda}=1+\frac{c(V+k)}{\Lambda V^2}$. Hence, $\frac{d\lambda^{FB}}{d\mu}\Big|_{\mu=\Lambda+c/V}=\frac{1}{1+\frac{c(V+k)}{\Lambda V^2}}=\frac{\Lambda V^2}{\Lambda V^2+c(V+k)}$. Therefore, $\Pi'(\Lambda+c/V)=\frac{\Lambda V^2}{\Lambda V^2+c(V+k)}-\omega(\Lambda+c/V)$. If $\frac{\Lambda V^2}{\Lambda V^2+c(V+k)}-\omega(\Lambda+c/V)>0$, i.e., $\omega<\frac{\Lambda V^2}{[\Lambda V^2+c(V+k)](\Lambda+c/V)}\triangleq\bar{\omega}$, then there exists a unique $\mu_0>\Lambda+c/V$ such that $\Pi'(\mu_0)=0$. Also, From (A.9), $d\mu/d\lambda^{FB}>1$ implies $d\lambda^{FB}/d\mu<1$. It follows from $d\lambda^{FB}/d\mu\,|_{\mu=\mu_0}-\omega\mu_0=0$ that $\mu_0<1/\omega$. If $\omega\geq\bar{\omega}$, then $\mu\leq\Lambda+c/V$ and the optimal capacity follows from the FIFO case. Thus, $\mu^{FB}=\mu_0>\Lambda+c/V$ (with $\mu^{FIFO}=\Lambda+c/V$) if $\omega<\bar{\omega}$ and $\mu^{FB}=\mu^{FFIO}$ otherwise. This completes the proof. □

Proof of Proposition 9 If $\omega\in[\bar{\omega},V/(2c))$, the first best is trivially achieved by non-referral FIFO. Next, consider $\omega<\bar{\omega}$. If $k\geq V-c/\mu_0$, then under μ_0, referrals cannot be generated (by Proposition 3). Thus, the firm runs a FIFO queue with $\mu^{\mathrm{SB}}=\Lambda+c/V<\mu_0$. If $k\geq V$, then obviously $k\geq V-c/\mu_0$ and then non-referral FIFO is optimal. This proves Case (iv) of the Proposition.

Next, we consider $k<V$. By Proposition 4, a referral mechanism must satisfy: $\max_{\mu,r}\ \frac{\Lambda}{1-r}-\frac{1}{2}\omega\mu^2$,

$$\text{s.t.}\quad V-rk-\frac{c}{\mu(1-r)-\Lambda}\geq 0, \tag{A.10}$$

$$V-k-\frac{c}{\mu(1-r)-\Lambda r}\geq 0, \tag{A.11}$$

whereas a FIFO mechanism (with all customers joining) has the optimal objective function value $\Pi^{FIFO}=\Lambda-\frac{1}{2}\omega(\Lambda+c/V)^2$. Note that since $\omega<\bar{\omega}$, the optimal FIFO mechanism must have all customers join the queue. Also note that in the referral mechanism formulation, both Constraints (A.10) and (A.11) being non-binding is not optimal because one can always increase r (without changing μ) to satisfy both constraints to strictly increase the objective function value. Thus, the referral mechanism can have three possibilities at optimality. (1) Both Constraints (A.10) and (A.11) are binding (which corresponds to full priority and it achieves the first best, which dominates the FIFO mechanism); (2) Constraint (A.10) is binding, but not Constraint (A.11) (which corresponds to partial priority and it achieves the first best, which dominates the FIFO mechanism); (3) Constraint (A.11) is binding, but not Constraint (A.10) (which corresponds to

strategic delay; it does not achieve the first best and we must also compare the optimal objective value with the FIFO mechanism to decide which one is better). In the first best (μ^{FB}, r^{FB}), since $w < \bar{\omega}$, $r^{FB} > 0$ and $V - r^{FB}k - \frac{c}{\mu^{FB}(1-r^{FB})-\Lambda} = 0$. We examine conditions under which (μ^{FB}, r^{FB}) satisfies (A.11) with $\mu = \mu^{FB}$ and $r = r^{FB}$. Note that as ω decreases, μ^{FB} increases and so does r^{FB}. From $V - r^{FB}k - \frac{c}{\mu^{FB}(1-r^{FB})-\Lambda} = 0$ we have $\mu^{FB}(1 - r^{FB}) = \Lambda + \frac{c}{V-kr^{FB}}$. Plugging it into the following gives: $V - k - \frac{c}{\mu^{FB}(1-r^{FB})-\Lambda r^{FB}} = V - k - \frac{c}{\Lambda + c/(V-kr^{FB}) - \Lambda r^{FB}} \triangleq \hbar(r^{FB})$. Recognize that $\hbar(1) = 0$. Also, let $h(r) \triangleq c/(V - kr) - \Lambda r$. Thus, $\hbar(r) = V - k - \frac{c}{\Lambda + h(r)}$. $h'(r) = \frac{ck}{(V-kr)^2} - \Lambda$. By inspection, $h'(r)$ is increasing in r. Therefore, $h(r)$ is convex in $r \in (0,1)$, and since $\hbar(r)$ is increasing in $h(r)$, $\hbar(r)$ is quasi-convex in $r \in (0,1)$. Since $\hbar(r)$ is quasi-convex and $\hbar(1) = 0$, we have the following three possibilities regarding the signs of $\hbar(0)$ and $\hbar'(1)$: Case 1: $\hbar(0) \geq 0$ and $\hbar'(1) \leq 0$. Case 2: $\hbar(0) \leq 0$ and $\hbar'(1) \geq 0$. Case 3: $\hbar(0) > 0$ and $\hbar'(1) > 0$.

Case 1: $\hbar(0) \geq 0$ and $\hbar'(1) \leq 0$. In this case, $\hbar(r) \geq 0, \forall r \in (0,1)$. Further, $\hbar'(1) \leq 0$ is equivalent to $h'(1) \leq 0$. Thus, $\hbar(0) = V - k - \frac{c}{\Lambda + c/V} \geq 0$, and $\hbar'(1) = \frac{ck}{(V-k)^2} - \Lambda \leq 0$. These conditions are equivalent to $\Lambda \geq \frac{ck}{(V-k)^2} \triangleq \bar{\Lambda}$. Hence, if $\Lambda \geq \bar{\Lambda}$, $\hbar(r) > 0$ for all any $r \in (0,1)$, which implies the first best can be achieved by partial priority for all $\omega < \bar{\omega}$. This proves Case (i) of the Proposition.

Case 2: $\hbar(0) \leq 0$ and $\hbar'(1) \geq 0$. In this case, $\hbar(r) < 0, \forall r \in (0,1)$. Conditions $\hbar(0) \leq 0$ and $\hbar'(1) \geq 0$ is equivalent to

$$V - k - \frac{c}{\Lambda + c/V} \leq 0. \tag{A.12}$$

That is, $\Lambda \leq \frac{ck}{(V-k)V} \triangleq \underline{\Lambda}$. In this case, the first best cannot be achieved for any $w < \bar{w}$. Thus, the second best (μ^{SB}, r^{SB}) is either FIFO, or a referral mechanism with strategic delay that satisfies $\chi(\mu^{SB}, r^{SB}) \triangleq V - k - \frac{c}{\mu(1-r^{SB})-\Lambda r^{SB}} = 0$, $V - r^{SB}k - \frac{c}{\mu(1-r^{SB})-\Lambda} > 0$. Note that when $\mu = \Lambda + c/V$ and $r = 0$, $\chi(\mu, r) = V - k - \frac{c}{\Lambda + c/V} < 0$. Also, $\chi(\mu, r)$ is increasing in μ and decreasing in r. Hence, any (μ, r) that satisfies $\chi(\mu, r) = 0$ must have $\mu > \Lambda + c/V = \mu^{FIFO}$. This implies that in this case, $\mu^{SB} > \mu^{FIFO}$. From $\chi(\mu, r) = 0$, we have

$$\mu = \frac{\Lambda r + c/(V-k)}{1-r} = \frac{\Lambda + \Lambda(r-1) + c/(V-k)}{1-r} = -\Lambda + \frac{\Lambda + c/(V-k)}{1-r}. \tag{A.13}$$

Let $x = \frac{\Lambda}{1-r}$. $\frac{\Lambda}{1-r} - \frac{1}{2}\omega\mu^2 = x - \frac{\omega}{2}\left[-\Lambda + x\left(1 + \frac{c}{\Lambda(V-k)}\right)\right]^2 \triangleq \Pi(x)$. $\Pi(x)$ is a concave quadratic function. The first-order condition is $\Pi'(x) = 1 + \Lambda\omega\left(1 + \frac{c}{\Lambda(V-k)}\right) - \omega x\left(1 + \frac{c}{\Lambda(V-k)}\right)^2$, $x \geq \Lambda$. $\Pi'(x)$ is decreasing in x and $\Pi'(\infty) = -\infty$. Thus, if $\Pi'(\Lambda) > 0$ then Π is maximized at x^*, where x^* is the unique solution to $\Pi'(x^*) = 0$; Otherwise, $\Pi(x)$ is decreasing in x, and the maximizer $x^* = \Lambda$ (i.e., no referrals). Also, $\Pi(\Lambda) = \Lambda - \frac{\omega}{2}\left(\frac{c}{V-k}\right)^2 < \Lambda - \frac{\omega}{2}\left(\Lambda + \frac{c}{V}\right)^2 = \Pi^{FIFO}$, where the inequality follows from (A.12). This implies that when $x^* = \Lambda$, the optimal second-best mechanism is non-referral FIFO. $\Pi'(\Lambda) = 1 - \omega\left(1 + \frac{c}{\Lambda(V-k)}\right)\frac{c}{(V-k)}$. $\Pi'(\Lambda)$ is decreasing in ω; thus, when ω is small enough, $\Pi'(\Lambda) > 0$ and there exists a unique $x^* > 0$ such that $\Pi'(x^*) = 0$. We next show that when $\omega = \bar{\omega}$, $\Pi'(\Lambda) < 0$. $\Pi'(\Lambda)\,|_{\omega=\bar{\omega}} = 1 - \bar{\omega}\left(1 + \frac{c}{\Lambda(V-k)}\right)\frac{c}{(V-k)} = 1 - \frac{\Lambda V^2}{[\Lambda V^2 + c(V+k)](\Lambda + c/V)}\left(1 + \frac{c}{\Lambda(V-k)}\right)\frac{c}{(V-k)}$. From (A.12), $c/(V-k) > \Lambda + c/V$. Hence, $1 - \frac{\Lambda V^2}{[\Lambda V^2 + c(V+k)](\Lambda + c/V)}\left(1 + \frac{c}{\Lambda(V-k)}\right)\frac{c}{(V-k)} < 1 - \frac{\Lambda V^2}{[\Lambda V^2 + c(V+k)](\Lambda + c/V)}\left(1 + \frac{c}{\Lambda(V-k)}\right)(\Lambda + c/V) = 1 - \frac{V^2[\Lambda(V-k)+c]}{[\Lambda V^2 + c(V+k)](V-k)} = 1 - \frac{\Lambda V^2(V-k) + cV^2}{\Lambda V^2(V-k) + c(V^2-k^2)} < 0$. Letting $\Pi'(\Lambda) = 0$ gives $w = \frac{\Lambda(V-k)^2}{c[\Lambda(V-k)+c]} \triangleq \tilde{\omega}$. Therefore, for $\omega \in [\tilde{\omega}, \bar{\omega})$, the optimal second-best mechanism is non-referral FIFO. When $\omega < \tilde{\omega}$, the referral mechanism

with strategic delay (SD) is feasible. Its system throughput is $\lambda^{SD} = \frac{1+\Lambda\omega\left(1+\frac{c}{\Lambda(V-k)}\right)}{\omega\left(1+\frac{c}{\Lambda(V-k)}\right)^2}$. The optimal capacity of the referral mechanism is $\mu^{SD} \triangleq \frac{\Lambda(V-k)}{\omega[\Lambda(V-k)+c]}$. The optimal profit under that mechanism is $\Pi^{SD} = \frac{\Lambda^2(V-k)(V-k+2(c+\Lambda(V-k))\omega)}{2(c+\Lambda(V-k))^2\omega} = \frac{\Lambda^2(V-k)^2}{2(c+\Lambda(V-k))^2\omega} + \frac{\Lambda^2(V-k)}{(c+\Lambda(V-k))}$. The difference between the referral profit and the FIFO profit $\Pi^{FIFO} = \Lambda - \frac{1}{2}\omega(\Lambda + c/V)^2$ is

$$\Pi^{SD} - \Pi^{FIFO} = \frac{\Lambda^2(V-k)^2}{2(c+\Lambda(V-k))^2\omega} + \frac{(\Lambda+c/V)^2}{2}\omega + \frac{\Lambda^2(V-k)}{(c+\Lambda(V-k))} - \Lambda. \tag{A.14}$$

Next, we prove that $\Pi^{SD} - \Pi^{FIFO}$ is monotone decreasing in ω for $\omega < \tilde{\omega}$. Differentiating $\Pi^{SD} - \Pi^{FIFO}$ with respect to ω gives $\Omega(\omega) \triangleq -\frac{\Lambda^2(V-k)^2}{2(c+\Lambda(V-k))^2\omega^2} + \frac{(\Lambda+c/V)^2}{2}$, which is increasing in ω. Hence, it suffices to show $\Omega(\tilde{\omega}) < 0$. $\Omega(\tilde{\omega}) = -\frac{c^2}{2(V-k)^2} + \frac{(\Lambda+c/V)^2}{2} < 0$, where the last inequality is due to (A.12). Since $\Pi^{SD} - \Pi^{FIFO}$ is monotone decreasing in ω for $\omega < \tilde{\omega}$ and $\Pi^{SD} - \Pi^{FIFO} > 0$ as $\omega \to 0$ and $\Pi^{SD} - \Pi^{FIFO} < 0$ when $\omega = \tilde{\omega}$, there exists a unique $\hat{\omega}$ such that $\Pi^{SD} - \Pi^{FIFO} > 0$ if and only if $\omega < \hat{\omega}$. Therefore, the optimal mechanism is strategic delay for $\omega < \hat{\omega}$, and non-referral FIFO for $\omega \geq \hat{\omega}$. This proves Case (iii) in the Proposition.

Next, we show that the optimal capacity under strategic delay in the second best, μ^{SD}, is lower than the optimal capacity in the first best, μ^{FB}. Let $\lambda_{FB}(\mu)$ and $\lambda_{SD}(\mu)$ be the throughput as a function of capacity μ in the first best and the second best (with strategic delay), respectively. Thus, μ^{FB} solves $\lambda_{FB}{}'(\mu^{FB}) - \omega\mu^{FB} = 0$ and μ^{SD} solves $\lambda'_{SD}(\mu^{SD}) - \omega\mu^{SD} = 0$. Note that both $\lambda'_{FB}(\mu)$ and $\lambda'_{SD}(\mu)$ are decreasing in μ. Therefore, to show $\mu^{SD} < \mu^{FB}$, it suffices to show $\lambda'_{FB}(\mu) > \lambda'_{SD}(\mu)$. From (A.9), $\frac{d\mu}{d\lambda_{FB}} = 1 + \frac{c\lambda_{FB}[\lambda_{FB}(V-k)+2k\Lambda]}{\Lambda[\lambda_{FB}(V-k)+k\Lambda]^2}$. From (A.13), $\frac{d\mu}{d\lambda_{SD}} = 1 + \frac{c}{\Lambda(V-k)}$. Since $dy/dx = 1/(dx/dy)$, to show $\lambda'_{FB}(\mu) > \lambda'_{SD}(\mu)$, it suffices to show $\frac{d\mu}{d\lambda_{FB}} < \frac{d\mu}{d\lambda_{SD}}$, i.e., $\frac{c\lambda_{FB}[\lambda_{FB}(V-k)+2k\Lambda]}{\Lambda[\lambda_{FB}(V-k)+k\Lambda]^2} < \frac{c}{\Lambda(V-k)}$. This is equivalent to showing $(V-k)\lambda_{FB}[\lambda_{FB}(V-k)+2k\Lambda] < [\lambda_{FB}(V-k)+k\Lambda]^2$, which is equivalently showing $(k\Lambda)^2 > 0$, which is trivially true. Hence, $\mu^{SD} < \mu^{FB}$.

Case 3: $\hbar(0) > 0$ and $\hbar'(1) > 0$. This gives the condition $\Lambda \in (\underline{\Lambda}, \bar{\Lambda})$. In this case, there exists a unique $r' \in (0,1)$ such that $\hbar(r) > 0$ for $r < r'$; $\hbar(r) = 0$ for $r = r'$; and $\hbar(r) < 0$ for $r > r'$. Since r^{FB} decreases in ω, it follows that there exists a unique $\underline{\omega}$ such that the first best is achieved by partial priority if $\omega \in (\underline{\omega}, \bar{\omega})$, by full priority if $\omega = \underline{\omega}$, and cannot be achieved in the second best if $\omega < \underline{\omega}$. Next, we show that strategic delay dominates FIFO for $\omega < \underline{\omega}$. First, at $\omega = \underline{\omega}$, $\Pi^{SD}(\underline{\omega}) = \Pi^{FB}(\underline{\omega}) > \Pi^{FIFO}(\underline{\omega})$ and $\mu^{SD}(\underline{\omega}) = \mu^{FB}(\underline{\omega}) > \mu^{FIFO}(\underline{\omega}) = \Lambda + c/V$. Second, from (A.14), $\Pi^{SD}(\omega) - \Pi^{FIFO}(\omega)$ is decreasing in ω if $\omega < \frac{\Lambda(V-k)}{[\Lambda(V-k)+c](\Lambda+c/V)}$. From the expression of μ^{SD}, we have $\mu^{SD}(\underline{\omega}) = \frac{\Lambda(V-k)}{\underline{\omega}[\Lambda(V-k)+c]}$. Since $\mu^{SD}(\underline{\omega}) > \Lambda + c/V$, we have $\underline{\omega} < \frac{\Lambda(V-k)}{[\Lambda(V-k)+c](\Lambda+c/V)}$. Therefore, $\Pi^{SD}(\omega) - \Pi^{FIFO}(\omega)$ is decreasing in ω if $\omega < \underline{\omega}$. Since $\Pi^{SD}(\underline{\omega}) > \Pi^{FIFO}(\underline{\omega})$, we have $\Pi^{SD}(\omega) > \Pi^{FIFO}(\omega)$ for all $\omega < \underline{\omega}$. Thus, strategic delay dominates FIFO and is optimal for $\omega < \underline{\omega}$. This proves Case (ii) of the Proposition. Note that $\mu^{SD} \in (\mu^{FIFO}, \mu^{FB})$ can be similarly shown as before.

Finally, when referrals are generated, the average delay $rW_1 + (1-r)W_0$ is equal to $(V - rk)/c$ because $V - rk - c[rW_1 + (1-r)W_0] = 0$. In a FIFO queue, the average delay W^{FIFO} is equal to V/c. Therefore, $rW_1 + (1-r)W_0 < W^{FIFO}$. This completes the proof. □

References

Mršević, M (2008) Convexity of the inverse function. The *Teaching of Mathematics* XI(1):21–24.

Appendix B: Equilibrium Conditions for the Full-Priority Referral Scheme

Under the full-priority referral scheme, if a customer makes a successful referral, she gains full priority over all the others who do not make a successful referral. Those with a successful referral (without a successful referral) expect a delay of W_1 (W_0). An equilibrium is specified by customers' joining probability $\alpha \in [0,1]$ and referral effort $r \in [0,1]$. In equilibrium, (W_1, W_2) is determined from (α, r) from the following equations: $W_1(r,\alpha) = \frac{1}{\mu(1-r\alpha)-\Lambda r\alpha^2},\ W_0(r,\alpha) = \frac{\mu(1-r\alpha)}{[\mu(1-r\alpha)-\Lambda r\alpha^2][\mu(1-r\alpha)-\Lambda\alpha]}$.

Case 1: $(\alpha<1, r=0)$ is an equilibrium if $V - \frac{1}{\mu-\Lambda\alpha} = 0,\ c\alpha\left(\frac{1}{\mu-\Lambda\alpha} - \frac{1}{\mu}\right) < k$.

Case 2: $(\alpha=1, r=0)$ is an equilibrium if $V - \frac{1}{\mu-\Lambda} \geq 0,\ c\left(\frac{1}{\mu-\Lambda} - \frac{1}{\mu}\right) < k$.

Case 3: $(\alpha<1, r\in(0,1))$ is an equilibrium if $V - kr - \frac{1}{\mu(1-r\alpha)-\Lambda\alpha} = 0,\ c\alpha\left[W_0(r,\alpha) - W_1(r,\alpha)\right] = k$.

Case 4: $(\alpha=1, r\in(0,1))$ is an equilibrium if $V - kr - \frac{1}{\mu(1-r)-\Lambda} \geq 0,\ c\left[W_0(r,1) - W_1(r,1)\right] = k$.

Case 5: $(\alpha<1, r=1)$ is an equilibrium if $V - k - \frac{1}{\mu(1-\alpha)-\Lambda\alpha} = 0,\ c\alpha\left[W_0(1,\alpha) - W_1(1,\alpha)\right] \geq k$.

Appendix C: Mechanism Design Formulation that Discriminates Between Base and Referred Customers

The mechanism design problem chooses expected delays $\boldsymbol{W} = (W_1^B, W_0^B, W_1^R, W_0^R)$ to induce customer strategies $\sigma \triangleq (\alpha^B, \alpha^R, r^B, r^R)$ that maximize the system throughput, subject to the IR, IC, OA and stability constraints. For ease of exposition, let $q = r^B\alpha^R,\ p = r^R\alpha^R, \lambda = \Lambda\alpha^B$. The mechanism design formulation is as follows. The objective function maximizes the system throughput: $\max_{\sigma,\boldsymbol{W}} \frac{\lambda(1+q-p)}{1-p}$. IR constraint 1: $V - kr^B - c[r^B\alpha^R W_1^B + (1-r^B\alpha^R)W_0^B] \geq 0$. IR constraint 2: $V - kr^R - c[r^R\alpha^R W_1^R + (1-r^R\alpha^R)W_0^R] \geq 0$. IC constraints: $r^B \in \arg\max_{r'} V - kr' - c[r'\alpha^R W_1^B + (1-r'\alpha^R)W_0^B],\ r^R \in \arg\max_{r'} V - kr' - c[r'\alpha^R W_1^R + (1-r'\alpha^R)W_0^R]$. Stability constraint: $\frac{\lambda(1+q-p)}{1-p} < \mu$. Next, we list fifteen (15) OA constraints. We invoke Lemma B.1 of Yang and Debo (2019): In a batch-arrival queue with exponential service rate μ, Poisson arrival rate λ, and random batch size N with finite first moment $\mathbb{E}[N]$ and second moment $\mathbb{E}[N^2]$, the expected delay W in the system is $W = \frac{\mathbb{E}[N]+\mathbb{E}[N^2]}{2\mathbb{E}[N](\mu-\lambda\mathbb{E}[N])}$. OA constraint 1: $W_1^B \geq \frac{1}{\mu-\lambda q}$. OA constraint 2: $W_0^B \geq \frac{1}{\mu-\lambda(1-q)}$. OA constraint 3: $W_1^R \geq \frac{1}{\mu(1-p)-\lambda qp}$. OA constraint 4: $W_0^R \geq \frac{1}{\mu-\lambda q}$. OA constraint 5: $qW_1^B + (1-q)W_0^B \geq \frac{1}{\mu-\lambda}$. OA constraint 6:

$$\lambda q W_1^B + \frac{\lambda qp}{1-p} W_1^R \geq \frac{\lambda q}{1-p}\frac{1}{\mu(1-p)-\lambda q}. \tag{C.1}$$

OA constraint 7: $\lambda q W_1^B + \lambda q W_0^R \geq 2\lambda q \frac{3}{2[\mu-2\lambda q]}$. OA constraint 8: $\frac{\lambda qp}{1-p}W_1^R + \lambda(1-q)W_0^B \geq \frac{\lambda(1-p-q+2pq)}{1-p}\frac{pq+(1-q)(1-p)^2}{(1-p-q+2pq)(\mu(1-p)-\lambda(1-p-q+2pq))}$. OA constraint 9: $\frac{\lambda qp}{1-p}W_1^R + \lambda q W_0^R \geq \frac{\lambda q}{1-p}\frac{1}{\mu(1-p)-\lambda q}$. OA constraint 10: $\lambda(1-q)W_0^B + \lambda q W_0^R \geq \frac{\lambda}{\mu-\lambda}$. OA constraint 11: $\lambda q W_1^B + \frac{\lambda qp}{1-p}W_1^R + \lambda(1-q)W_0^B \geq \frac{\lambda(1-p+pq)}{1-p}\frac{(1-q)(1-p)^2+q}{(1-p+pq)[\mu(1-p)-\lambda(1-p+pq)]}$. OA constraint 12: $\lambda q W_1^B + \frac{\lambda qp}{1-p}W_1^R + \lambda q W_0^R \geq \frac{\lambda q(2-p)}{1-p}\frac{(1-p)^2+2}{2(2-p)[\mu(1-p)-\lambda(2-p)]}$. OA constraint 13: $\frac{\lambda qp}{1-p}W_1^R + \lambda(1-q)W_0^B + \lambda q W_0^R \geq \frac{\lambda(1-p+pq)}{1-p}\frac{(1-q)(1-p)^2+q}{(1-p+pq)[\mu(1-p)-\lambda(1-p+pq)]}$. OA constraint 14: $\lambda q W_1^B + \lambda(1-q)W_0^B + \lambda q W_0^R \geq \lambda(1+q)\frac{1+2q}{(1+q)(\mu-\lambda(1+q))}$. OA constraint 15:

$$\lambda q W_1^B + \frac{\lambda qp}{1-p}W_1^R + \lambda(1-q)W_0^B + \lambda q W_0^R \geq \frac{\lambda(1+q-p)}{1-p}\frac{1+q-p+(1-p)(q-p)}{(1+q-p)[\mu(1-p)-\lambda(1+q-p)]}. \tag{C.2}$$

In terms of the structure of the optimal referral mechanism, partial priority (subject to work conservation) corresponds to (C.1) being non-binding and (C.2) being binding; strategic delay (with full priority assigned to referring customers) corresponds to (C.1) being binding and (C.2) being non-binding; full priority (subject to work conservation) corresponds to both (C.1) and (C.2) being binding; non-referral FIFO corresponds to $r^B = r^R = 0$.