Invite Your Friend and You'll Move up in Line: Leveraging Social Ties via Operational Incentives

Luyi Yang
Booth School of Business, The University of Chicago, luyi.yang@chicagobooth.edu
Laurens Debo
Tuck School of Business, Dartmouth College, laurens.g.debo@tuck.dartmouth.edu

The referral priority program—an emerging business practice adopted by a growing number of technology companies that manage a waitlist of customers—enables existing customers on the waitlist to gain priority access if they successfully refer new customers to the waitlist. Unlike more commonly used referral reward programs, this novel mechanism does not offer monetary compensation to referring customers, but leverages customers' own disutility of delays to create referral incentives. Despite this appealing feature, our queueing-game-theoretic analysis finds the effectiveness of such a scheme as a marketing tool for customer acquisition and an operational approach for waitlist management depends crucially on the underlying market conditions, particularly the base market size of spontaneous customers. The referral priority program might not generate referrals when the base market size is either too large or too small. When customers do refer, the program could actually backfire, namely, by reducing the system throughput and customer welfare, if the base market size is intermediately large. This phenomenon occurs because the presence of referred customers severely cannibalizes the demand of spontaneous customers. We also compare the referral priority program with the referral reward program when the service provider optimally sets the admission price. We find that under a small base market size, the referral reward program would encourage referrals using monetary incentives. Numerical study suggests the referral priority program is more profitable than the referral reward program when the base market size is intermediately small.

Key words: referrals; priority queues; system throughput; customer welfare; conversion rate; batch arrivals

1. Introduction

Many technology companies are breaking new ground today as they introduce sign-up waitlists for a limited release to eager customers before making their products available to the general public. Notable examples include Dropbox, a file-hosting service (Ries 2011), Mailbox, an email inbox-management application (Shontell 2013), and Robinhood, a mobile application for commission-free stock trading (Roberts 2015). These companies have enjoyed sensational success in attracting an inundation of waitlist sign-ups. Despite a variety of reasons for using a waitlist (see, e.g., Hamburger 2013), many would agree the primary motivation is to clear technological hurdles and validate beta products with real users to "ensure that everyone has a fantastic, reliable experience" (in the words

of Robinhood). Thus, it behooves these firms to take customers off the waitlists at a steady yet limited rate, which may cause excessive wait times.

Recognizing this situation, many firms embrace a novel mechanism that allows customers to move up in line if they invite their friends to also sign up on the waitlist. For instance, Robinhood's confirmation email reads, "Interested in priority access? Get early access by referring your friends." Referrals have become such an integral part of a waitlist that, for example, Waitlisted.co—a startup specialized in helping client companies build waitlists—has made it a standard built-in feature to "spread word of mouth by allowing users to improve their queue position by referring people." We call this emerging business practice the *referral priority program*.

The ingenuity of the referral priority program is that it cleverly leverages customers' dislike of waiting to create an incentive for spreading positive word of mouth and acquiring new customers on behalf of the firm. Compared to the traditional referral reward program, which offers monetary compensation to motivate referrals, the referral priority program "recruits" existing customers as sales agents without the firm incurring any explicit costs or proactively designing the reward payments. Integrating such a free and hands-off referral program into a waitlist holds immense appeal, especially for firms whose tight budget constraints prohibit the use of monetary rewards.

Another key characteristic of the referral priority program is that it takes advantage of not only social ties between existing customers and their friends (as in all other referral programs), but also of interactions among customers on the waitlist. A customer's spot in line is relative, and non-referring customers could move backwards when referring customers are granted priority access. Thus, the amount of priority one obtains with a successful referral depends on others' referral behavior. Moreover, as referrals bring in new customers, the system could suffer more congestion, which may, in turn, diminish customers' willingness to sign up. If a customer anticipates her friend is unlikely to convert due to such congestion, she might also choose not to refer. Therefore, customers' incentives in this setting are intricate and warrant closer scrutiny.

This paper focuses on customers' strategic joining and referral behavior, and investigates whether and when the program benefits the firm as a marketing tool and customers as an operational choice. To model customer referrals on a waitlist, we consider a queueing game played by delay-sensitive customers. *Base customers* arrive to the queueing system (waitlist) spontaneously at a rate termed the *base market size*. Customers make both joining and referral decisions based on their rational beliefs of expected delays in different priority classes and the probability that a *referred customer* converts, i.e., joins the queue. Customers who make a successful referral join the priority class, and customers who do not refer or refer in vain are placed in a regular class. Our assumption of instantaneous referral conversion gives rise to a tractable priority queue with batch arrivals. In equilibrium, referred customers are less likely to join and more likely to refer than base customers.

Referred customers expect a longer delay in either priority class, which makes them less likely to join. On the other hand, the relative delay difference between the priority class and the regular one is more significant for referred customers, increasing their referral incentives.

We find that referrals are generated when the base market size is intermediate and the customer population sufficiently values the service. If the base market size is too small, the benefit of gaining priority is too incremental to cover the cost of referrals, because congestion is light in the first place. Thus, customers would rather not refer. If the base market size is too large, despite a strong incentive for priority, the conversion rate of referrals is low because referred customers are turned away by excessive congestion. Anticipating that referrals are most likely to be futile, base customers would choose not to refer. Moreover, a higher service valuation by the customer population is conducive to referrals, because it would allure more joining, thereby increasing the conversion rate and stimulating the need for priority.

When referrals are generated, the sign-ups of referred customers *cannibalize* the demand of base customers. More base customers balk because their expected delay is prolonged by the presence of referred customers. Thus, how the system throughput would fare becomes unclear. If the base market size is intermediately small (but not small enough to completely discourage referrals), we show that the demand creation of referred customers is the primary force and the referral program is effective in enhancing the system throughput. Nevertheless, if the base market size is intermediately large (but not large enough to deter referrals altogether), the demand cannibalization effect becomes so severe that the system throughput will actually be *lower* than if the referral priority program is not used. Intuitively, when the base market size is relatively large, the conversion rate is necessarily low, and thus the additional arrivals of referred customers would not compensate for the loss of base customers. In addition, this adverse phenomenon would be partially countervailed when the customer population places a higher valuation on the service.

When the referral priority program harms the system throughput, we find it always reduces customer welfare. Even when the referral program expands the market, customer welfare may still be lower. Hence, although peer-to-peer referrals bring value to those who would be unaware of the service otherwise and thus not join (referred customers), they may not always justify the increased system congestion and the loss of base customers.

We also consider the firm's optimal pricing in the referral priority program and compare it with the traditional referral reward program in which the firm optimally determines both the price for admission to the waitlist and a monetary reward to compensate referrals. We find that contrary to the referral priority program, which precludes referrals when the base market size is small, the optimal referral reward program would, in fact, motivate referrals under a small base market size.

Our numerical comparison shows the referral priority program is more profitable than the referral reward program when the base market size is intermediately small.

Our results have important managerial implications for firms that entertain the use of referral priority programs within their waitlists as a means to acquire new customers. We show that while the referral priority program may outperform the referral reward program in some cases, it may hurt companies operating under some other market conditions. Thus, firms cannot be agnostic of their business environment when deciding whether to run the referral priority program, but must be serious with their market research, especially in terms of gauging the base market size.

The remainder of this paper is organized as follows. §2 reviews related literature. In §3, we set up the main model of the joining-referral game among customers in the referral priority program. §4 characterizes the equilibrium of the game. In §5, we examine when customers would refer, and when the referral priority program would be (in)effective in improving system throughput and customer welfare. §6 studies the firm's optimal pricing problem in the referral priority program, and compares the optimal profit between the referral priority program and the referral reward program. §7 concludes the paper with a summary of our main results, and discusses future research.

2. Related Literature

Our research connects the marketing literature on word-of-mouth/customer referrals and the operations literature on customers' strategic behavior in queueing systems.

2.1. Word of Mouth and Customer Referrals

Word of mouth communication and consumer social interaction have been well recognized as important factors in designing marketing strategies (Buttle 1998, Godes et al. 2005). Most relevant to our research is the growing body of literature that tackles the design of referral reward programs in which customers are compensated with a monetary reward upon successful referrals.

The seminal paper by Biyalogorsky et al. (2001) considers how the firm should jointly set the purchase price and referral reward. They find the referral reward program alleviates the free-riding problem caused by a low price due to its "pay for performance" nature. Kornish and Li (2010) design the optimal referral bonuses when referrals not only disseminate product information, but also signal product quality. Keeping price fixed, Xiao et al. (2011) investigate how to provide two-way incentives in referral reward programs to both referring and referred customers. Lobel et al. (2016) study the impact of the social network structure in designing referral payment when firms value referrals but can only compensate conversions. Libai et al. (2003) take up the problem of setting referral fees in a related setting of affiliate marketing, in which merchants contract affiliates for inducting customers to their websites. Jing and Xie (2011) compare the referral reward program with group buying, another selling mechanism based on social interaction. They find group buying

is superior in achieving a larger scale of social interaction, whereas the referral reward program is more effective in discriminating between customers according to their referral outcome.

Our paper contributes to this rich body of theoretical literature on referral reward programs by studying a new mechanism in which the "reward" is priority access on the waitlist. As such, the amount customers receive is no longer a decision variable controlled by the firm, but rather an equilibrium outcome determined by customers' self-interested referral behavior.

Experimental and empirical works have documented the positive value of referral programs. Schmitt et al. (2011) find referred customers have a higher contribution margin, a higher retention rate, and are more valuable in both the short run and long run. Garnefeld et al. (2013) show participation in referral programs also reinforces referring customers' loyalty. Additionally, evidence suggests customer satisfaction, deal proneness, and tie strength (Wirtz and Chew 2002), as well as brand strength, and the recipient of rewards (Ryu and Feick 2007) may affect the effectiveness of referral rewards. Skepticism about incentivized referrals (e.g., Trusov et al. 2009, Verlegh et al. 2013) usually revolves around the potential distortion created by monetary rewards, arguing such referrals may be less cost effective to the firm offering rewards and less trustworthy to customers who receive referral links. Our paper sheds light on a different source of concern in incentivized customer referrals when monetary rewards are not involved: instead of expanding the market, the referral priority program may actually dampen demand.

2.2. Strategic Behaviors in Queues

This literature dates back to Naor (1969), in which customers decide whether to join or balk after observing the queue length. Edelson and Hildebrand (1975) consider this joining versus balking problem in unobservable queues. We refer the reader to Hassin and Haviv (2003) for an extensive survey. Our paper is particularly related to research on strategic behavior in *priority* queues.

Kleinrock (1967) proposes a bidding mechanism in which customers can bribe the service provider for priority. Lui (1985) and Glazer and Hassin (1986) pin down customers' priority bidding behavior in Nash equilibrium. Hassin (1995) shows such a decentralized priority auction maximizes social welfare. Follow-up works consider various extensions, including a generalized delay cost structure (Afèche and Mendelson 2004) and private information on job-processing time (Kittsteiner and Moldovanu 2005). Other papers examine an alternative setting where the service provider posts prices for different priority classes. Mendelson and Whang (1990) find the socially optimal priority prices that are also incentive compatible for customers whose waiting costs follow a discrete distribution. Afèche (2013) study this problem from a revenue-maximizing perspective and show inserting strategic delay might be optimal. Gavirneni and Kulkarni (2016) investigate a problem in which customers with continuously distributed waiting costs can self-select into a priority class by paying an extra fee. All these papers examine unobservable queues as in ours.

A smaller stream of literature studies priority purchasing in observable queues. Balachandran (1972) characterizes stable payment policies as a function of the queue length. In Adiri and Yechiali (1974), customers who observe the system state can pay a fee for priority. They show customers follow strategies of control-limit type, and purchase priority when the queue length is above a certain threshold. Hassin and Haviv (1997) demonstrate the "follow the crowd" behavior in Adiri and Yechiali (1974) can lead to multiple and potentially mixed-strategy equilibria. Alperstein (1988) find the welfare-maximizing pricing policy in the setting of Adiri and Yechiali (1974) would implement a last-come, first-served queuing discipline.

Common in all the above papers is that priority can be purchased with a premium price. Our paper contributes three novel distinctions to this stream of literature. First, in priority purchasing schemes, priority is guaranteed with a premium price, but in referral priority programs, whether one obtains priority is probabilistic, depending on referred customers' endogenous joining decision. Second, because of this stochastic conversion, our model of unobservable queues does not rely on customers' ex-ante heterogeneity in waiting costs to generate two priority classes (unlike most unobservable models); instead, we identify a novel type of ex-post heterogeneity in terms of the source of customers, i.e., arrivals out of spontaneity or from referrals. Third, priority prices are internal transfers among customers and the service provider, which only affects the *arrival rate* of customers, whereas referrals would change the *arrival process* as new customers are brought in. Specifically, referrals in our model give rise to a priority queue with batch arrivals.

Thus, on a technical side, our paper builds on the queueing literature of batch arrivals. Burke (1975) studies a single-class batch arrivals queue without priorities. Hawkes (1965) combines batch arrivals with priority queues and assumes customers in each batch all join the same priority class. The queueing system in our model of the referral priority program is most similar to the ones studied in Takahashi and Takagi (1990) and Takagi and Takahashi (1991), where customers in a given batch could join different priority classes. The reader is referred to Chaudhry and Templeton (1983) for a survey of batch-queueing models. The literature on strategic behavior in queues with batch arrivals is relatively scant. Yildirim and Hasenbein (2010) consider the admission control and pricing problem in a batch-arrivals queue in which each arriving batch collectively makes joining and balking decisions. Ziani et al. (2015) study a Markovian queue with batch arrivals of two customers who individually decide whether to join or balk. In these papers, the batch size or its probability distribution is exogenously specified, whereas in our model, the batch forms endogenously due to customers' priority-incentivized referrals.

3. Model

We model the sign-up waitlist as a single-server queueing system. The service time is i.i.d. exponentially distributed with mean $1/\mu$. *Base customers*, denoted by "B," arrive to the system according

to a Poisson process with rate Λ. These customers are aware of the service and arrive spontaneously (not from referrals). We call Λ the base market size.

All customers have a common waiting cost per unit time $c > 0$ and valuation for service v drawn from a uniform distribution over $[0, \bar{V}]$. Each arriving customer decides whether to join the queue, and upon joining, each customer may make one referral. *Referred customers*, denoted by "R," arrive instantaneously upon receiving referral requests, and also decide on joining and referring. The same process continues for a friend's friend and so on. Customers incur a referral cost $c_r \geq 0$ if they invite a friend. If a referral is successful, i.e., the referred customer joins, the referring customer joins the priority class (Class 1). If a customer does not refer or if her referral is unsuccessful, a customer joins the regular class (Class 2). Class 1 customers are served under preemptive priority over class 2 customers. Within each class, customers are served FIFO (First In, First Out).

We assume customers do not observe the queue length of either class. The model primitives, Λ, μ, c, c_r and the valuation distribution (including $\bar{V}$) are common knowledge. A customer's information set consists of her valuation and type (whether she is a base customer or a referred one), (v, χ), where $v \in [0, \bar{V}], \chi \in \{B, R\}$. That is, (i) an individual customer's valuation for service v is private information not known by other customers; (ii) a customer knows her own type, but a referred customer does not know the type of her referrer, or of her referrer's referrer and so on. Of course, by definition, a friend any customer refers is a referred customer.

Customers make rational joining-referral decisions at the time of arrivals. To decide whether to join or refer, a customer must form beliefs about the expected delay she is subject to in each of the priority classes and the probability that her friend joins (because she does not observe her friend's valuation). Upon arrival, each customer chooses a pair of actions (j, r), where $j \in \{0, 1\}$ indicates whether the customer joins ($j = 1$ for joining), and $r \in \{0, 1\}$ indicates whether the customer refers ($r = 1$ for referring). The action space is $A = \{(0, 0), (1, 0), (1, 1)\}$. A pure strategy is a mapping $\sigma(v, \chi) : [0, \bar{V}] \times \{B, R\} \mapsto A$, specifying an action given customer valuation v and customer type χ (base or referred). Let the strategy space be Σ.

We consider symmetric equilibria of the joining-referral game among customers. Given everyone else's strategy σ, let $W_i^{\chi}(\sigma)$ be the induced expected delay (including time spent at service) in Class $i = 1, 2$ for customer type $\chi \in \{B, R\}$, and let $\alpha(\sigma) \in [0, 1]$ be the conversion rate induced by σ, or the probability that a referred customer joins. If everyone else plays strategy σ, a customer playing strategy $\sigma'(v, \chi) = (j', r'), \forall (v, \chi) \in [0, \bar{V}] \times \{B, R\}$ yields an expected utility:

$$U_{\sigma', \sigma}(v, \chi) = \begin{cases} 0, & (j', r') = (0, 0) \\ v - cW_2^{\chi}(\sigma), & (j', r') = (1, 0) \\ v - c_r - c[\alpha(\sigma) W_1^{\chi}(\sigma) + (1 - \alpha(\sigma)) W_2^{\chi}(\sigma)], & (j', r') = (1, 1) \end{cases}. \tag{1}$$

If a customer does not join, her utility is normalized to zero. If a customer joins and does not refer, she joins Class 2, expects a delay $W_2^\chi(\sigma)$, and, therefore, her expected utility is service value v less the expected waiting cost $cW_2^\chi(\sigma)$. If a customer joins and refers, with probability $\alpha(\sigma)$, her friend joins, and she advances to Class 1 that has expected delay $W_1^\chi(\sigma)$; with probability $1-\alpha(\sigma)$, her friend does not join, and she still joins Class 2 with expected delay $W_2^\chi(\sigma)$. Hence, the expected utility for a customer who joins and refers is service value v less referral cost c_r and the expected waiting cost $c[\alpha(\sigma)W_1^\chi(\sigma)+(1-\alpha(\sigma))W_2^\chi(\sigma)]$.

In a Bayes Nash equilibrium, customers form beliefs over the expected delay and the response of their friends (in terms of the conversion rate). Given these beliefs, they choose actions to maximize their expected utility, and these actions result in the expected delay and conversion rate consistent with initial beliefs. A pure symmetric-strategy Bayes Nash equilibrium $\sigma \in \Sigma$ satisfies:

$$U_{\sigma,\sigma}(v,\chi) \geq U_{\sigma',\sigma}(v,\chi), \quad \forall (v,\chi) \in [0,\bar{V}] \times \{B,R\}, \sigma' \in \Sigma.$$

Given other customers' strategy, the optimal referral action for a joining customer of type χ is independent of v. Hence, customer χ joins if and only if her valuation is weakly above a certain cutoff value v^χ. That is, a random base customer joins with probability $\beta \triangleq 1-\mathbb{P}(v \geq v^B)$; a random referred customer joins with probability $\alpha \triangleq 1-\mathbb{P}(v \geq v^R)$. Because customer valuation is uniformly distributed over $[0,\bar{V}]$, we have $v^B = \bar{V}(1-\beta),\ v^R = \bar{V}(1-\alpha)$. Thus, an equilibrium is more conveniently characterized by a tuple $\boldsymbol{s} = (\beta, \alpha, r^B, r^R)$, where $\beta, \alpha \in [0,1]$ are just as defined, and $r^B, r^R \in \{0,1\}$ dictate whether joining base and referred customers refer, respectively. More generally, customers could play a mixed referral strategy. This possibility can be easily accommodated by modifying the interpretation of $r^\chi \in [0,1]$ to be the probability that customer χ refers a friend, $\chi \in \{B,R\}$.

3.1. Discussion of the Model

Because customers make rational joining and balking decisions, taking into account the impact of the referral program, the resulting queueing system will be stable in equilibrium. All referred customers joining and continuing to refer cannot be sustained in equilibrium because the system would be too crowded, which, in turn, leads referred customers to balk. On the other hand, a customer may have no friends to refer; alternatively, an invited friend may not be interested in the service, or is inclined to ignore any referral requests received, or may already be on the waitlist. We can incorporate these possibilities by superimposing an exogenous dampening factor $\gamma < 1$ on the conversion rate. Adding this additional parameter would have no material impact on the model.

Allowing for one referral per customer at most gives rise to a parsimonious model of two priority classes. The marketing literature (e.g., Biyalogorsky et al. 2001, Kornish and Li 2010) routinely

adopts the same single-referral assumption. Likewise, the operations literature often focuses on two priority classes in queueing systems to glean insights (e.g., Hassin and Haviv 1997, Afèche 2013). With a single referral, customers spread the word in a tractable "chain" structure: a batch-arrivals queue forms with batch size following a modified geometric distribution (as we will see in §4.1). By contrast, with multiple referrals, a "tree" structure emerges: the arrival batch would become a branching process, with customers grouped into different priority classes by the number of successful referrals they make, i.e., the offsprings they produce. The batch size is the total number of individuals ever born in this branching process before the population becomes extinct. Customers' decision problems are also more subtle, because a customer may be later overtaken by the very friend she refers if the friend makes more successful referrals.

In essence, referrals necessitate correlation among customer arrivals. Our assumption that referred customers arrive instantaneously results in a batch-queueing model (as we will see in §4.1) as the simplest means to elegantly capture such demand correlation. A model incorporating lead time in the response of a friend may lack tractability because one needs to keep track of the customers who have made a referral but not yet received a response. Even when the lead time is exponentially distributed, the arrival process would be a modification of the Yule-Furry process (e.g., Ross 1996, pp. 235) with immigration (due to base arrivals), leading to a queueing system that does not admit simple closed-form expressions of expected delays. The game-theoretic analysis is also more involved. Customers may choose not to refer if they expect to be served before their friends respond. Crossovers in referral conversions are possible, i.e., a referring customer may be overtaken by future arriving customers while waiting for the referral response.

Most of the applications that motivate our model are virtual queues, and, therefore, whether customers observe the queue length is left to the firm's delay-announcement policy. For instance, LBRY, a decentralized content-sharing and publishing platform, does not provide real-time delay information within its referral priority program. On the other hand, Robinhood discloses the aggregate queue-length information, yet does not report to customers the size of each priority class. We assume the queue is unobservable mainly for analytical tractability, and embedding the firm's delay-announcement policy in the modeling framework of the referral priority program is an interesting question beyond the scope of the present paper.

4. Equilibrium

In this section, we characterize the equilibria of the referral-joining game. First, in §4.1, we derive the expressions for the expected delay in Class 1 and 2 for base and referred customers, respectively, under a given equilibrium conjecture $\boldsymbol{s} = (\beta, \alpha, r^B, r^R)$. We then specify the equilibrium conditions in §4.2 using the expressions derived in §4.1.

4.1. Queueing Preliminaries

In the presence of the referral priority program, customer arrivals no longer follow a Poisson process in general, but rather a compound Poisson process with batch arrivals. A batch forms at the instant of a base customer's arrival, because a sequence of referred customers would come successively until a customer stops referring or fails to refer successfully. Thus, given the equilibrium conjecture $\boldsymbol{s}$, the effective arrival rate is $\Lambda\beta$; the batch size N is a random variable following a modified geometric distribution:

$$\mathbb{P}(N=1)=1-q, \quad \mathbb{P}(N=k)=q(1-p)p^{k-2}, k\geq 2, \qquad \mathbb{E}[N]=\frac{1+q-p}{1-p}, \tag{2}$$

where $q=r^B\alpha$ is the unconditional probability that a joining base customer brings in a friend, which occurs if the base customer refers (with probability r^B) and the referral recipient joins (with probability α); and $p=r^R\alpha$ is the unconditional probability that a joining referred customer brings in a friend, which occurs if the referred customer refers (with probability r^R) and the referral recipient joins (with probability α). The *system throughput* is thus $\Lambda\beta(1+r^B\alpha-r^R\alpha)/(1-r^R\alpha)$.

Notice that here, because all customers who refer successfully join Class 1, the first $N-1$ customers in the batch join Class 1, and the last customer in the batch joins Class 2. Due to batch arrivals, the expected delay in a given class for an arriving customer is not equal to the average delay of that class from a system's standpoint. In other words, the PASTA (Poisson Arrivals See Time Average) property (Wolff 1982) does not directly apply. Moreover, a base customer (the first one in a batch) would expect a different delay than referred customers (subsequent ones in a batch) in joining a given priority class, which is why our model stipulates that a customer's information set include her type. Here, the implicit (yet reasonable) assumption is that customers are aware of whether they arrive spontaneously or from referrals, but referred customers do not possess information about their own positions in the batch.

Now we extract the essence of the underlying queueing system. It is a single-server preemptive priority queue with exponentially distributed service time with mean $1/\mu$, a Poisson arrival rate λ, and an arrival batch size following a modified geometric distribution as in (2); if the total batch size is N, the first $N-1$ customers in the batch join the priority class (Class 1), and the last customer joins the regular class (Class 2). We refer to this queueing system as a priority queue with batch arrivals. System stability requires $\mu>\lambda(1+q-p)/(1-p)$. We operate under this assumption on the parameters[1]. Let ω_1^B and ω_2^B be the expected delay for the first customer in the batch (base customer) if she joins Class 1 and Class 2, respectively. Let ω_1^R and ω_2^R be the expected delay for a customer who is not the first in the batch (referred customer) if she joins Class 1 and Class 2, respectively. Lemma 1 gives closed-form expressions of these expected delays.

[1] In our referral-joining game, stability will always be endogenously satisfied in equilibrium due to customers' rational joining behavior.

LEMMA 1. *In the priority queue with batch arrivals,*

$$\omega_1^B(\lambda,q,p)=\frac{\mu(1-p)^2+\lambda pq}{\mu(1-p)(\mu(1-p)-\lambda q)},\quad \omega_1^R(\lambda,p,q)=\omega_1^B(\lambda,q,p)+\frac{1}{\mu(1-p)},$$
$$\omega_2^B(\lambda,q,p)=\frac{\mu(1-p)^2+\lambda q}{[\mu(1-p)-\lambda q][\mu(1-p)-\lambda(1+q-p)]},\quad \omega_2^R(\lambda,q,p)=\omega_2^B(\lambda,q,p)+\frac{1}{\mu(1-p)-\lambda q}.$$

It is immediate from Lemma 1 that referred customers expect a longer delay than base customers in a given priority class. This result is intuitive because the base customer is the first in the batch. Also, it is easy to verify from Lemma 1 that a referred customer's expected delay is at least $2/\mu$ regardless of which class she joins, because her delay includes at least her own service time plus the service time of the customer who refers her (both are $1/\mu$ on average). The fact that a referred customer becomes aware of the service, by definition, implies that her referring customer makes a successful referral and joins Class 1. The referred customer has to at least wait behind her. To exclude trivial cases with no scope for referred customers to join, we introduce Assumption 1.

ASSUMPTION 1. $\bar{V}>2c/\mu$.

4.2. Equilibrium Referral Strategies

Equipped with the expected delay expressions in Lemma 1 for a given equilibrium conjecture, we now turn to the characterization of equilibrium. We start with customers' best response in referrals. From the utility functions in (1), customer χ refers (given others' strategy σ) if and only if

$$v-c_r-c[\alpha(\sigma)W_1^\chi(\sigma)+(1-\alpha(\sigma))W_2^\chi(\sigma)]\geq v-cW_2^\chi(\sigma).$$

Rewriting terms yields

$$c_r\leq c\alpha(\sigma)[W_2^\chi(\sigma)-W_1^\chi(\sigma)]. \tag{3}$$

The left-hand side (LHS) of (3) is the cost of referring a friend, and the right-hand side (RHS) of (3) is the expected benefit of a referral. It shows that the referral incentive is driven by two factors: (i) the incentive to gain priority, determined by the relative difference in expected delays of two classes, and (ii) the likelihood that a friend converts. Factor (ii) highlights the distinction between the incentive to refer and the incentive to gain priority. If the conversion rate α is low, referrals may not be justified even when joining the priority class confers a substantial delay reduction.

COROLLARY 1. *For any λ,p,q satisfying $\mu>\lambda(1+q-p)/(1-p)$, we have $\omega_2^B(\lambda,q,p)-\omega_1^B(\lambda,q,p)\leq\omega_2^R(\lambda,q,p)-\omega_1^R(\lambda,q,p)$ with equality at $q=0$. Therefore, in equilibrium, $r^B\leq r^R$ with equality at $r^B=r^R=0$ or $r^B=r^R=1$.*

Corollary 1 directly follows from Lemma 1 and suggests that given any equilibrium conjecture, referred customers expect a larger difference in expected delays of the two priority classes than

base customers. Thus, from (3), it follows that referred customers have a greater incentive to refer against any equilibrium conjecture. The intuition is as follows. From Lemma 1, referred customers would expect a longer delay in either class than base customers. Their longer delay in the regular class will be further exacerbated because more time spent there engenders more chances of being overtaken by future customers. Hence, referred customers are worse off than base customers in joining the regular class relative to the priority class. Referred customers' greater incentive to refer rules out the possibility of equilibria in which base customers refer while referred customers do not. It is also impossible that both base and referred customers play a mixed referral strategy because doing so would imply these two types of customers enjoy the same expected benefit of referrals (equal to c_r). Thus, we are left with four possible forms of referral strategies:

(i) $(r^B, r^R) = (0,0)$: neither base nor referred customers refer.

(ii) $(r^B, r^R) = (1,1)$: both base and referred customers refer.

(iii) $(r^B, r^R) = (0,\kappa),\ \kappa \in (0,1)$: base customers do not refer; referred customers randomize.

(iv) $(r^B, r^R) = (\kappa,1),\ \kappa \in (0,1)$: base customers randomize; referred customers refer.

Initially, considering referral strategy (iii) might seem odd. If base customers do not refer, referred customers do not even exist and their referral strategy seems irrelevant. However, a non-referring base customer must also assess her expected utility from referring a friend, which depends on her belief about the conversion rate α, which, in turn, hinges on referred customers' referral/joining strategy. Viewed in the light of a sequential game framework, base customers are the first mover and referred customers can only act in a subgame. Even if a subgame is not reached, the "off-the-equilibrium-path" behavior is still important because it disciplines the behavior in equilibrium.

4.2.1. Strict Non-referral Equilibrium: $(r^B, r^R) = (0,0)$. Neither base nor referred customers refer. The equilibrium (β, α) satisfies the following conditions:

$$\bar{V}(1-\beta) - \frac{c}{\mu - \Lambda\beta} = 0, \tag{4a}$$

$$c_r \geq c\alpha \left(\frac{1}{\mu - \Lambda\beta} - \frac{1}{\mu} \right), \tag{4b}$$

$$\text{If } \ \bar{V} - c\left[\frac{1}{\mu - \Lambda\beta} + \frac{1}{\mu} \right] < 0,\ \alpha = 0; \ \text{ otherwise, } \ \bar{V}(1-\alpha) - c\left[\frac{1}{\mu - \Lambda\beta} + \frac{1}{\mu} \right] = 0. \tag{4c}$$

This strict non-referral equilibrium results in an $M/M/1$ queue with expected delay $1/(\mu - \Lambda\beta)$ and throughput $\Lambda\beta$. Equation (4a) pins down β by setting the expected utility of the "marginal customer" to zero. $\bar{V}(1-\beta) = v^B$ is the marginal base customer's valuation for service. Condition (4b) indicates base customers prefer not to invite a friend. Because no one else refers, a successful referral would move a customer to the head of the queue with an expected delay $1/\mu$ (just the service time). The conversion rate α in Condition (4b) is specified by hypothesizing referred customers'

joining and referral behavior off the equilibrium path. Because of base customers' non-referrals ($q = r^B\alpha = 0$), according to Corollary 1, (fictitious) referred customers would expect the same delay difference in the two priority classes, and thus would also not refer. Hence, referred customers' expected delay is $1/(\mu - \Lambda\beta) + 1/\mu$, where the term $1/\mu$ accounts for the additional delay caused by waiting behind the referring base customer. Condition (4c) determines α in much the same way as equation (4a): $\bar{V}(1-\alpha) = v^R$ is the marginal referred customer's valuation for service.

Because the LHS of equation (4a) is decreasing in β, we can uniquely pin down β. Plugging this β into Condition (4c) uniquely determines α. Thus, (4b) is a verification condition for the equilibrium. If it is satisfied, a unique strict non-referral equilibrium exists. We call this system the "FIFO benchmark," and denote the equilibrium β by β^F, and the equilibrium throughput by λ^F.

4.2.2. All-Referral Equilibrium: $(r^B, r^R) = (1, 1)$. Both base and referred customers refer. The equilibrium (β, α) satisfies the following conditions:

$$\bar{V}(1-\beta) - c_r - c[\alpha W_1^B(\alpha,\beta) + (1-\alpha)W_2^B(\alpha,\beta)] = 0, \tag{5a}$$

$$c_r \le c\alpha\left[W_2^B(\alpha,\beta) - W_1^B(\alpha,\beta)\right], \tag{5b}$$

$$\bar{V}(1-\alpha) - c_r - c[\alpha W_1^R(\alpha,\beta) + (1-\alpha)W_2^R(\alpha,\beta)] = 0, \tag{5c}$$

where $W_i^\chi(\alpha,\beta) = \omega_i^\chi(\Lambda\beta, \alpha, \alpha),\ i = 1, 2,\ \chi \in \{B, R\}$. In this all-referral equilibrium, the throughput is $\Lambda\beta/(1-\alpha)$. With a slight abuse of notation, we express W_i^χ as a function of α and β. Conditions (5a) and (5c) jointly determine α and β by requiring that marginal base and referred customers have zero expected utility, assuming they refer. Condition (5b) guarantees that the expected benefit of referrals is large enough to cover the referral cost for base customers. This condition implies referred customers would also refer, because the expected benefit of referrals for them is even larger.

4.2.3. Weak Non-referral Equilibrium: $(r^B, r^R) = (0, \kappa)$. Base customers do not refer; referred customers randomize. The equilibrium (β, α, κ) satisfies the following conditions:

$$\bar{V}(1-\beta) - c/(\mu - \Lambda\beta) = 0, \tag{6a}$$

$$c_r = c\alpha\left(\frac{1}{\mu - \Lambda\beta} - \frac{1}{\mu}\right), \tag{6b}$$

$$\bar{V}(1-\alpha) - c\left[\frac{1}{\mu - \Lambda\beta} + \frac{1}{\mu(1-\kappa\alpha)}\right] = 0. \tag{6c}$$

In this weak non-referral equilibrium, the throughput is $\Lambda\beta$, where β is determined by equation (6a), which is the same as equation (4a), and hence $\beta = \beta^F$. This equilibrium also implements the same $M/M/1$ FIFO queue as the strict non-referral equilibrium, yet under different conditions, as shown in (6b) and (6c). Equation (6b) ensures (fictitious) referred customers are indifferent to referrals, which supports their mixed referral strategy. In fact, by Corollary (1), base customers here

($q = r^B\alpha = 0$) have the same referral incentive as referred customers and thus are also indifferent to referrals, but choose not to refer (to sustain this equilibrium). We can solve for α from equation (6b) by plugging β from equation (6a). Condition (6c) determines the randomization probability $\kappa \in (0,1)$ by setting the marginal referred customer's expected utility to zero if she does not refer and expects a delay $1/(\mu - \Lambda\beta) + 1/(\mu(1-\kappa\alpha))$. Because referred customers are indifferent to referrals and thus randomize, we can equivalently specify Condition (6c) using the case in which they refer. Similar to the strict non-referral equilibrium, the weak non-referral equilibrium (β, α, κ) can be uniquely determined (if it exists).

4.2.4. Partial-Referral Equilibrium: $(r^B, r^R) = (\kappa, 1)$**.** Base customers randomize; referred customers refer. The equilibrium (β, α, κ) satisfies the following conditions:

$$\bar{V}(1-\beta) - c_r - c[\alpha W_1^B(\alpha,\beta,\kappa) + (1-\alpha)W_2^B(\alpha,\beta,\kappa)] = 0, \tag{7a}$$

$$c_r = c\alpha[W_2^B(\alpha,\beta,\kappa) - W_1^B(\alpha,\beta,\kappa)], \tag{7b}$$

$$\bar{V}(1-\alpha) - c_r - c[\alpha W_1^R(\alpha,\beta,\kappa) + (1-\alpha)W_2^R(\alpha,\beta,\kappa)] = 0, \tag{7c}$$

where $W_i^\chi(\alpha,\beta) = \omega_i^\chi(\Lambda\beta, \kappa\alpha, \alpha), \quad i = 1,2,\ \chi \in \{B, R\}$.

In this partial-referral equilibrium, the throughput is $\Lambda\beta\,[1 + \kappa\alpha/(1-\alpha)]$. The equilibrium values of (β, α, κ) are jointly determined by equations (7a) through (7c). Conditions (7a) and (7c) require that marginal base and referred customers have zero expected utility if they choose to refer. Condition (7b) guarantees the expected benefit of referrals is equal to the referral cost for base customers, and therefore, base customers are indifferent to referrals. This indifference condition implies that referred customers strictly prefer to invite a friend because they have a stronger incentive to refer.

Given the model primitives, if a tuple $(\beta, \alpha, r^B, r^R)$ satisfies any of the four sets of equilibrium conditions above, it constitutes an equilibrium.

4.3. Existence of Equilibria and Structural Results

We first define a cutoff market size $\bar{\Lambda}$ as follows:

$$\bar{\Lambda} \equiv \mu \frac{\bar{V}(\bar{V} - 2c/\mu)}{(\bar{V} - c/\mu)c/\mu}. \tag{8}$$

One can show by straightforward algebra that $\bar{\Lambda}$ is increasing in $\bar{V}$. Building on the definition of $\bar{\Lambda}$, Proposition 1 establishes structural properties of the all-referral equilibrium.

PROPOSITION 1. *If and only if $\Lambda < \bar{\Lambda}$, there exists a unique cutoff value $c_r^l > 0$ such that for referral cost $c_r \in [0, c_r^l]$, there exists a unique all-referral equilibrium $(r^B, r^R) = (1,1)$ and when $c_r = c_r^l$, base customers are indifferent to referrals. Furthermore, the equilibrium joining probabilities of base and referred customers (β, α) and the system throughput are all decreasing in $c_r \in [0, c_r^l]$.*

Results in Proposition 1 are consistent with intuition. All customers refer if the referral cost is sufficiently small. Moreover, further decreasing the referral cost makes joining more attractive for both base and referred customers, and therefore, the equilibrium β and α increases, leading to a higher throughput. This equilibrium would arise under some c_r if the base market size Λ is smaller than $\bar{\Lambda}$, where $\bar{\Lambda}$ is the cutoff base market size solved for by setting α and c_r to zero in equations (5a) and (5c). If Λ is too large, the expected delay in the system would be too overwhelming for referred customers to even join, and referrals would be futile due to non-conversion. Thus, the all-referral equilibrium would no longer be sustained. We note that for $c_r \in [0, c_r^l]$, although there exists a unique all-referral equilibrium, equilibria of other forms may also exist. Leveraging the structural result in Proposition 1, we formally establish the existence of equilibria in Theorem 1.

THEOREM 1. *There always exists an equilibrium in the form of one of the four possible referral strategies. Specifically, there exists c_r^l, c_r^m, c_r^h such that*

- $(r^B, r^R) = (0,0)$ *only if* $c_r \geq c_r^h$;
- $(r^B, r^R) = (0,\kappa)$ *only if* $c_r \in [c_r^m, c_r^h]$;
- $(r^B, r^R) = (1,1)$ *only if* $c_r \in [0, c_r^l]$;
- $(r^B, r^R) = (\kappa,1)$ *if* $c_r \in [\min\{c_r^l, c_r^m\}, \max\{c_r^l, c_r^m\}]$.

Either $c_r^l = c_r^m = c_r^h = 0$, *or* $c_r^l > 0$ *and* $c_r^h > c_r^m > 0$.

Theorem 1 shows the form of the equilibrium referral strategies crucially depends on the magnitude of the referral cost c_r. Intuitively, the smaller the referral cost, the more inclined customers are to refer. This intuition largely holds, as shown by different segments of c_r corresponding to different equilibrium forms in Theorem 1, but these segments may overlap, suggesting the possibility of multiple equilibria of different forms. Except for the pair $(0,0)$ and $(0,\kappa)$ (the two forms of non-referral equilibria), all other pairs of the four equilibrium forms can coexist for a given set of model primitives. Sometimes, there may even exist multiple equilibria of three different forms. Multiple equilibria are an artifact of the follow-the-crowd (FTC) behavior (see, e.g., Hassin and Haviv 1997): if others start to refer, a given customer would also be prompted to refer so as to keep her waiting position from being bumped.

Combining Proposition 1 and Theorem 1, we recognize that if $\Lambda \geq \bar{\Lambda}$, then $c_r^l = c_r^m = c_r^h = 0$, in which case customers do not make referrals under any referral cost.

COROLLARY 2. *If the referral cost is zero ($c_r = 0$), there exists a unique pure strategy equilibrium.*

Corollary 2 follows from Proposition 1 and Theorem 1. While we cannot guarantee uniqueness in general, at least it is always the case when referrals are free. Furthermore, in this case, either all customers refer or no customers refer.

COROLLARY 3. *In any equilibrium, the joining probability of referred customers is lower than that of base ones ($\alpha < \beta$).*

Corollary 3 points out that a random referred customer is less likely to join than base customers, whereas conditional on joining, referred customers are more eager to make referrals (Corollary 1). This demonstrates distinct characteristics of joining and referral incentives. Joining is driven by the *absolute* magnitude of the expected delays. The larger the delay, the less likely one would join. By contrast, referrals are driven by the *relative* difference between the expected delays in the two classes. The larger the difference, the more likely one would refer.

5. Effectiveness of the Referral Priority Program

The focus of this section is to study when the referral priority program is effective in acquiring more customers/increasing throughput as marketing tool. Specifically, we examine when customers would refer, when referrals would have a positive effect on throughput, and when it may have unintended consequences. We also investigate the program's welfare implications for customers.

5.1. Two Illustrative Examples

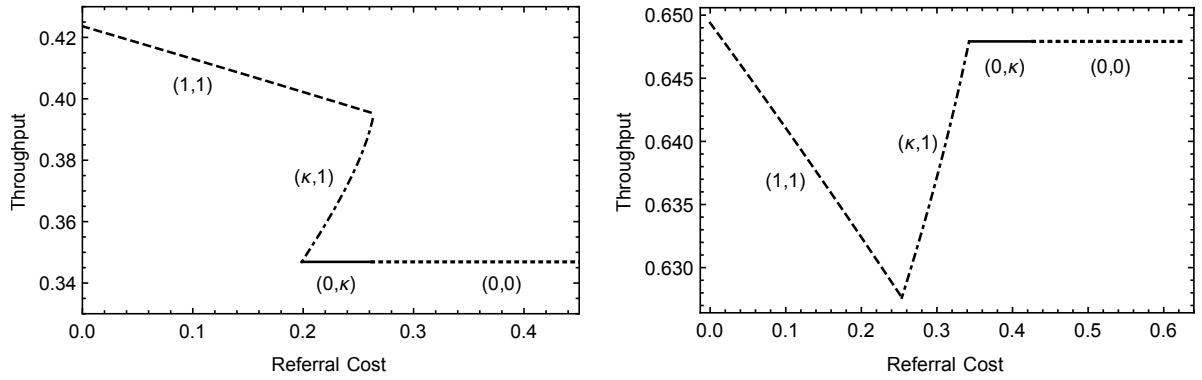
Figure 1 plots the equilibrium throughput achieved by the referral priority program against referral cost c_r. First, it illustrates that customers tend to engage in referrals when the referral cost gets small. Second, under small enough referral cost c_r where the all-referral equilibrium (1,1) appears, the system throughput rises as c_r falls, which illustrates the analytical result in Proposition 1. Third, when $\Lambda = 0.5$ (Figure 1-(a)), there may exist multiple equilibria for a given c_r when c_r is around 0.2 to 0.25. Yet, in any equilibrium that generates referrals, the resulting throughput is higher than the FIFO benchmark λ^F, represented by the flat line in the figure. By comparison, when $\Lambda = 1.5$ (Figure 1-(b)), there is always a unique equilibrium for a given c_r. However, for most c_r's under which referrals are generated, a *lower* throughput ensues (except when c_r is close to zero). This outcome is an unintended consequence of the referral priority program. The aim of encouraging referrals is to boost growth and facilitate customer acquisition. However, this example shows that referrals do not always translate to a higher throughput. We formalize the main observations from these two examples in our analytical development.

5.2. Analytical Results

We first study when the referral priority program is effective in generating referrals, i.e., inducing either the all-referral or partial-referral equilibrium.

PROPOSITION 2. *The referral priority program sustains an equilibrium that generates referrals if $\bar{V}$ is sufficiently high and Λ is intermediate; that is, there exist $\tilde{V}, \tilde{\Lambda}_l, \tilde{\Lambda}_h$ such that $r^R \geq r^B > 0$ is sustained in equilibrium if $\bar{V} > \tilde{V}$ and $\Lambda \in (\tilde{\Lambda}_l, \tilde{\Lambda}_h)$, where $\tilde{\Lambda}_h$ is increasing, and $\tilde{\Lambda}_l$ is decreasing in*

Figure 1 **Throughput against referral cost c_r under different base market sizes Λ.**



(a) $\Lambda = 0.5$. (b) $\Lambda = 1.5$.

Note. $\bar{V} = 5$, $c = 1$, $\mu = 1$. In the figure, $(1,1)$ represents the segments for all-referral equilibrium; $(\kappa, 1)$, partial-referral equilibrium; $(0, \kappa)$, weak non-referral equilibrium; $(0,0)$, strict non-referral equilibrium.

$\bar{V}$ for $\bar{V} > \tilde{V}$. In particular, if $c_r = 0$, then the referral priority program generates referrals if and only if $\Lambda \in (0, \bar{\Lambda})$.

Proposition 2 shows that if customers generally have a low valuation for service (the uniform distribution with a higher $\bar{V}$ stochastically dominates the one with a lower $\bar{V}$), customers do not refer regardless of the base market size. If the customer population has a high enough valuation for service, customers refer as long as the base market size Λ is intermediate. If the base market size is too low, the benefit of gaining priority is incremental because there is little congestion in the first place. Thus, customers would refrain from referrals. Following this logic, one would expect a larger base market size to be always conducive to referrals. Nevertheless, when the base market size is too large, significant balking kicks in as referred customers are turned away by excessive congestion. Therefore, a low conversion rate would, again, diminish the incentive to refer, despite a strong incentive for gaining priority. In the extreme case when referrals are free, the lower bound on Λ vanishes: referrals will be generated even if the base market size is arbitrarily small, but the upper bound ($\bar{\Lambda}$ as in equation (8)) persists. This result ties back to our discussion of Proposition 1: a base market size that is too large prevents referred customers from joining.

Moreover, Proposition 2 shows that with higher customer valuation for service, referrals would be generated under a wider range of base market sizes. Higher valuation encourages more customers to join, increasing both the conversion rate and the attractiveness of gaining priority. Thus, the referral disincentive either due to a large referral cost or an extreme base market size could be countervailed by higher valuation of the customer population.

Now, we turn to examine the equilibrium outcome when referrals are generated in a referral priority program, and compare it with the FIFO benchmark. We denote base customers' joining probability and the system throughput in those equilibria by β^R and λ^R, respectively.

PROPOSITION 3. *In any all-referral or partial-referral equilibrium, base customers join with a lower probability than they would under FIFO ($\beta^R < \beta^F$).*

Proposition 3 reveals that in the referral priority program, potentially creating demand from referred customers comes at the expense of cannibalizing demand from base customers who would otherwise join in the absence of the program. These base customers balk because their expected delay in the system is prolonged by the referral priority program. This phenomenon looks somewhat paradoxical considering that the referral program provides customers with an extra option, which should make joining more valuable. However, as more customers take up this option by bringing in their friends, the system becomes more congested than before, which, in turn, makes joining less attractive with this additional option. To this end, customers are "obliged" to refer not so much because they desire a shorter delay than they would get under FIFO, but rather because they simply wish to avoid the even longer delay in the regular class.

This result exposes a potential misconception about the referral priority program: it does not merely acquire more customers for free, but rather changes the mix of customers that adopt the service. On one hand, some base customers are lost (the effective arrival rate is lower). On the other hand, new customers are brought to the system (the batch size may be more than 1). These two opposing forces make it unclear in which direction the throughput would change. Next, we derive easily verifiable conditions on the model primitives to address this question.

THEOREM 2. *The referral priority program induces an equilibrium that reduces the throughput relative to FIFO ($\lambda^R < \lambda^F$) only if $\Lambda > \mu/2$.*

Theorem 2 provides a necessary condition under which the referral priority program could backfire. In other words, if the base market size is reasonably small relative to capacity ($\Lambda \leq \mu/2$), then the referral program may either not generate any referrals (and thus retain the FIFO throughput) or generate referrals that strictly increase the throughput relative to FIFO. Hence, the upward pressure on the throughput from acquiring referred customers always outweighs the downward pressure from losing base customers. Theorem 2 analytically confirms the observation from Figure 1-(a) that the referral priority program would never reduce the throughput when $\Lambda = 0.5$ and $\mu = 1$.

THEOREM 3. *If $\bar{V} > 5c/(2\mu)$ and $\Lambda \in \left(\underline{\Lambda}, \bar{\Lambda}\right)$, where $\underline{\Lambda} = \mu(3\bar{V})/[2(\bar{V} + 2c/\mu)]$, then there exists an interval $[c_{r,1}, c_{r,2}]$ such that for referral cost $c_r \in [c_{r,1}, c_{r,2}]$, the referral priority program induces an equilibrium that reduces the throughput relative to FIFO ($\lambda^R < \lambda^F$).*

Theorem 3 complements Theorem 2 by providing a sufficient condition under which the referral priority program could backfire. It indicates that when the base market size is sufficiently large but not large enough to dissuade referrals altogether (recall from Proposition 1 that $\bar{\Lambda}$ is the

upper bound on the base market size to generate referrals), implementing the referral priority program may be detrimental to the system throughput, i.e., the downward pressure from the balking of base customers overshadows the upward pressure from the joining of referred customers. This phenomenon could occur in either a partial-referral equilibrium or an all-referral equilibrium. Intuitively, an intermediately large base market size generally implies more congestion, which lowers joining probabilities and induces a relatively low conversion rate of referred customers. When referrals are triggered in this situation, referred customers' thin demand does not make up for the loss of base customers, and therefore, referrals would harm the overall system throughput.

We make three comments on the required conditions. First, $\bar{V} > 5c/(2\mu)$ ensures $\underline{\Lambda} < \bar{\Lambda}$. Otherwise, we would not find a Λ under which Theorem 3 holds. Specifically, when $\bar{V} = 5c/(2\mu)$, $\underline{\Lambda} = \bar{\Lambda} = 5\mu/6$. Second, it is easy to check that $\underline{\Lambda}$ is increasing in $\bar{V}$. Recall that $\bar{\Lambda}$ is also increasing in $\bar{V}$, which implies that if more customers in the population have a higher valuation for service, not only would customers be more prone to referrals, but the system throughput as a result of these referrals is also more likely to be higher than that under FIFO. Third, $\underline{\Lambda}$ asymptotically approaches $3\mu/2$ as $\bar{V}$ gets large. By comparison, $\bar{\Lambda}$ tends to infinity as $\bar{V}$ grows. Thus, a simple corollary is that for any $\Lambda \geq 3\mu/2$ and $\bar{V} > 3$ (which guarantees $\bar{\Lambda} > 3\mu/2$), implementing the referral priority program would cause the throughput to decline under some c_r. This result provides analytical support for the observation from Figure 1-(b) that the referral priority program may reduce the throughput when $\Lambda = 1.5$ and $\mu = 1$. This result also highlights the difference in the role customer population's valuation plays in driving referrals and in improving the throughput. While a high enough service valuation by the customer population can always prompt customers to refer, it only plays a modest role in offsetting the pernicious effect of a large base market size.

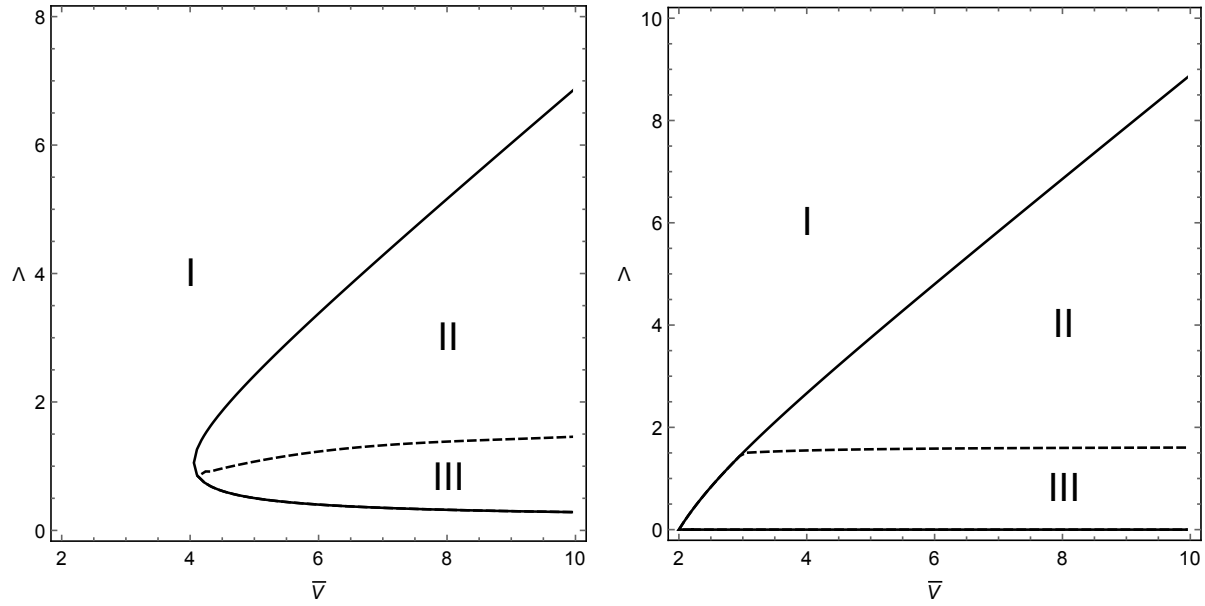
PROPOSITION 4. *For any $\bar{V} > 3c/\mu$, there exists a sufficiently small $\epsilon > 0$ such that under $\bar{\Lambda} - \epsilon$ and $c_r = 0$, the referral priority program reduces the throughput relative to FIFO ($\lambda^R < \lambda^F$).*

One may wonder to what extent the frictions present in making a referral (referral cost) impact our results. Intuitively, costly referrals make joining less attractive, and thus drive away too many customers. Indeed, Proposition 1 suggests the all-referral equilibrium achieves the highest throughput when $c_r = 0$. However, Proposition 4 shows that even when referrals are free, the referral priority program may still reduce the throughput, at a large base market size. Note that if the base market size is marginally below $\bar{\Lambda}$, an all-referral equilibrium emerges under $c_r = 0$ (see Proposition 1). We require $\bar{V} > 3c/\mu$ such that $\bar{\Lambda}$ would be relatively large. In other words, if $\bar{V} \leq 3c/\mu$, referrals would always yield a higher throughput than FIFO. Given the referral cost may be partially influenced by the firm (e.g., providing easily accessible referral links through multiple social media

channels), one implication of this result is that by making referrals easier, the firm might partially circumvent the decline in throughput, but would generally not eradicate the problem.

Figure 2 encapsulates much of the insights from the analytical results in this section. Fixing parameters c, μ, c_r, we can partition the $(\bar{V}, \Lambda)$ space into three regions when evaluating the effectiveness of the referral priority program. In Region I, the program does not generate referrals and maintains the FIFO throughput. Region II represents a particularly pessimistic scenario: the program generates referrals but reduces the throughput relative to the FIFO system in the absence of the program. Region III is the only region in which the referral priority program is effective: it generates referrals and increases the throughput relative to FIFO.

Figure 2 **Effectiveness of the referral priority program. The $(\bar{V}, \Lambda)$ can be partitioned into three regions. Region I: no referrals, the same throughput as FIFO; Region II: referrals generated, lower throughput than FIFO; region III: referrals generated, higher throughput than FIFO.**



(a) $c_r = 0.2$. (b) $c_r = 0$.

Note. $c = 1$, $\mu = 1$. With $c_r = 0$ in panel (b), equilibrium uniqueness is guaranteed due to Corollary 2. In panel (a), there may exist multiple equilibria, and the one with the lowest throughput is chosen to generate the plot. Numerically, we only observe multiple equilibria when Λ is small (around the boundary of Regions I and III). Other equilibrium selection criteria would generally not change the qualitative characteristics of the plot.

As shown in Figure 2-(a), when customer population's valuation for the service is low (manifested by a small $\bar{V}$), customers do not generate referrals under any base market size. When the customer population's valuation is relatively large, customers' referral decision depends on the base market size. Customers do not refer if the base market size is either too small or too large. When the

base market size is intermediate, customers generate referrals, but the resulting throughput will be higher than FIFO only when the base market size is intermediately small.

Note in Figure 2-(a) that a higher valuation by the customer population tends to expand the referral region (combining Region II and III) and the region in which referrals increase throughput (Region III). Whereas a higher base market size would usually induce a lower conversion rate, a higher service valuation by the customer population would entice more customers and thus induce a higher conversion rate, countervailing the referral disincentive or throughput decline caused by a large arrival base market size. Therefore, the referral priority program would be effective in boosting the system throughput when the customer population has high valuation toward the service, and when the base market size is intermediately small (fixing other parameters).

Figure 2-(b) shows that the same insights largely persist when referrals are costless. One main qualitative difference is that referrals will be generated even if the base market size is arbitrarily small (the part of Region I that is below Region III vanishes). This phenomenon is to be expected because customers no longer require large enough delay reduction to justify the referral cost. Figure 2-(b) also illustrates Proposition 4: if the environment is sufficiently close to the upper boundary of Region II, the throughput under the referral priority program is always strictly lower than that under FIFO (subject to $\bar{V} > 5c/(2\mu)$).

Next, we study the welfare implications of introducing the referral priority program. Denote the expected utility of a joining base customer with valuation v under equilibrium $\boldsymbol{s} = (\beta, \alpha, r^B, r^R)$ by

$$u^B(v, \boldsymbol{s}) = v - r^B \left(c\alpha W_1^B(\boldsymbol{s}) + c(1-\alpha) W_2^B(\boldsymbol{s}) + c_r \right) - \left(1 - r^B\right) c W_2^B(\boldsymbol{s}).$$

Denote the expected utility of a joining referred customer with valuation v under equilibrium $\boldsymbol{s}$ by

$$u^R(v, \boldsymbol{s}) = v - r^R \left(c\alpha W_1^R(\boldsymbol{s}) + c(1-\alpha) W_2^R(\boldsymbol{s}) + c_r \right) - \left(1 - r^R\right) c W_2^R(\boldsymbol{s}).$$

Since balking customers' utility is normalized to zero, we define total customer welfare:

$$CW = \Lambda\beta \left\{ \int_{\bar{V}(1-\beta)}^{\bar{V}} u^B(v, \boldsymbol{s}) \frac{1}{\bar{V}} \mathrm{d}v + \frac{r^B \alpha}{1-\alpha} \int_{\bar{V}(1-\alpha)}^{\bar{V}} u^R(v, \boldsymbol{s}) \frac{1}{\bar{V}} \mathrm{d}v \right\}.$$

Define individual customer welfare:

$$ICW = \frac{CW}{\Lambda\beta \left[1 + \frac{r^B \alpha}{1-\alpha} \right]}.$$

PROPOSITION 5. *Customer welfare has the following properties:*

(i) Individual customer welfare under the referral priority program is always weakly lower than that under FIFO;

(ii) If the referral priority program' throughput is lower than that under FIFO, so is its total customer welfare.

The result in Proposition 5 is surprising in the sense that customers are presented with an additional option (refer to gain priority); yet having this option may only make them worse off. Once referrals are generated, individual customer welfare is always lower than FIFO. This result is a consequence of demand cannibalization (Proposition 3) and weaker joining incentives of referred customers (Corollary 3). Demand cannibalization suggests joining is less attractive to base customers, and therefore, the expected utility of an average base customer is lower. Because referred customers have even lower joining incentives, the expected utility of an average referred customer is thus also lower than what an average customer would gain from a FIFO system. These two forces together imply lower individual customer welfare. As a result, in which direction total customer welfare moves is unclear, depending also on whether the market is expanded, i.e., whether the system throughput is increased. Thus, the referral priority program have even more difficulty improving total customer welfare than increasing throughput.

Self-interested customers ignore two sources of externalities when making referral decisions. The first is the commonly acknowledged negative externalities imposed on other customers because bringing friends to the system increases the amount of congestion (A referring customer partially internalizes the extra waiting costs imposed on the customers she overtakes by incurring a referral cost). The second is positive externalities in bringing value to new customers who would not otherwise join the system. If the benefit of expanding the market and creating value for more customers outweighs the increased waiting costs due to congestion plus the referral costs incurred, the referral priority program would be welfare-improving.

6. Optimal Pricing, Referral Reward Program, and Comparison

In previous sections, we focused on the analysis of the referral priority program itself, assuming customers obtain free access to the waitlist. In this section, we study the firm's optimal pricing decision as a monopolist to maximize the expected profit when it runs the referral priority program. Congestion pricing is an important lever to regulate the demand rate (Naor 1969, Edelson and Hildebrand 1975), whereas referrals alter the demand process. Thus, how pricing and referrals interact in shaping the firm's demand and profitability is an interesting question.

Moreover, incorporating optimal pricing provides a fair basis for profit comparison between the referral priority program and the referral *reward* program (see, e.g., Biyalogorsky et al. 2001) in which the firm motivates word of mouth by offering a monetary reward for a successful referral. The referral reward program is a *centralized* scheme in that the firm can control the arrival rate by the admission price, and the arrival process by the referral reward. By contrast, the referral priority

program is *decentralized* in that the firm only has the admission price lever, but cannot directly control whether customers refer. Given the complementary strengths of these two programs—the referral reward program has one more lever, whereas the referral priority program involves no explicit costs—the profit comparison of the two programs is less than obvious. §6.1 examines the optimal pricing problem in the referral priority program. In §6.2, we introduce the referral reward program and lay out the firm's joint optimization problem of the admission price and referral reward. §6.3 reports a numerical study that compares the two programs' profit performance; we will show the percentage change of profit in both programs relative to the optimal FIFO (non-referral) benchmark as in (9a)-(9b):

$$\max_{P\geq 0,\ \beta\in[0,1]} \quad P\Lambda\beta \tag{9a}$$

$$\text{s.t.} \quad \bar{V}(1-\beta) - P - \frac{c}{\mu - \Lambda\beta} = 0, \tag{9b}$$

where P and $\Lambda\beta$ is the optimal monopoly price and the concomitant system throughput, respectively, in the FIFO benchmark.

6.1. Pricing in the Referral Priority Program

Recall from §4.2 that customers have four possible referral strategies. The price the firm sets influences customers' referral strategies, because a higher price would imply a lower service valuation, which diminishes referral incentives (see Proposition 2). The firm does not know a priori which of the four referral strategies would be induced under the optimal price. Therefore, the firm's problem is to solve four profit-maximization problems, each corresponding to one particular referral-strategy conjecture, and then choose the one that yields the highest expected profit. The optimal solution to that chosen problem gives the optimal price. Following the convention of the literature (e.g. Mendelson and Whang 1990), we assume that given the optimal price, customers play the equilibrium that maximizes the firm's profit (if there ever exist multiple equilibria given the optimal price). We show the detailed formulation of the four optimization problems in Appendix B. We denote the referral priority program with optimal pricing by the optimal referral priority program.

Proposition 6. *The optimal referral priority program generates referrals only when the base market size Λ is intermediate.*

Proposition 6 is an extension of Proposition 2. Pricing provides the firm with a lever to control congestion as well as the admission of base customers and referred customers (conversion rate). When the base market size is too small, the firm must charge a low enough price to create a sufficient amount of congestion for customers to have an incentive to gain priority and thus generate referrals. If the price required is too low, the firm would rather forsake referrals. On the other hand, if the base market size is too large, the firm must also charge a low enough price to improve the conversion rate and consequently motivate referrals, which again may not be worthwhile.

6.2. Referral Reward Program

The firm charges price P for access to the service. Customers are served according to FIFO. In the referral reward program, no priority is credited toward referrals, but each customer who successfully brings in a friend receives a reward $\Delta \geq 0$ from the firm. Given admission price P and referral reward Δ, customers make joining and referral decisions. For fair comparison, we adopt the same modeling assumption about referrals as in the referral priority program (e.g., at most one referral, the referred customer arrives immediately, etc). Here, because the referral incentive is disentangled from the waiting incentive, customers refer if $\Delta\alpha \geq c_r$, where α is the endogenous conversion rate, and c_r is the referral cost. It immediately follows that the referral reward program's expected profit is at least as high as that of the optimal FIFO benchmark (9a)-(9b) because the firm can always set $\Delta = 0$ to shut down referrals. The same cannot be said about the referral priority program, because the firm does not directly control referrals.

Given admission price P and referral reward Δ, the equilibrium can be characterized by (β, α, κ), where, as before, β and α are the probability that a base and referred customer joins, respectively, and κ is the probability that a customer refers. Recall from the referral priority program that in general, base customers may follow a different referral strategy than referred customers. However, in the referral reward program, they would not because both types of customers are incentivized by the same reward Δ. Hence, κ applies to both types of customers. On the other hand, they would still adopt different joining strategies because their expected delays remain different. Specifically, referred customers would expect a longer expected delay than base ones (as we will see in Lemma 2). Let $W^{\chi}(\alpha, \beta, \kappa), \chi \in \{B, R\}$ be the expected delays for base and referred customers, respectively, for a given equilibrium conjecture (β, α, κ). The resulting queueing system is a batch-arrivals queue with arrival rate $\Lambda\beta$ and batch size following a geometric distribution with success probability $\kappa\alpha$. The first customer in the batch is a base customer, and the rest are referred customers.

LEMMA 2. *In the referral reward program for a given equilibrium conjecture* (β, α, κ)*:*

$$W^B(\alpha, \beta, \kappa) = \frac{\mu(1-\kappa\alpha)^2 + \Lambda\beta\kappa\alpha}{\mu\,(1-\kappa\alpha)\,[\mu(1-\kappa\alpha) - \Lambda\beta]}, \quad W^R(\alpha, \beta, \kappa) = W^B(\alpha, \beta, \kappa) + \frac{1}{\mu(1-\kappa\alpha)}.$$

Therefore, the firm's problem of setting the optimal price P and referral reward Δ is

$$\max_{P\geq 0, \Delta \geq 0;\ (\alpha,\beta,\kappa)\in[0,1]^3} \quad P\Lambda\beta + (P-\Delta)\frac{\Lambda\beta\kappa\alpha}{1-\kappa\alpha} \tag{10a}$$

$$\text{s.t.} \quad \bar{V}(1-\beta) - P - cW^B(\alpha, \beta, \kappa) + \kappa(-c_r + \alpha\Delta) = 0, \tag{10b}$$

$$\bar{V}(1-\alpha) - P - cW^R(\alpha, \beta, \kappa) + \kappa(-c_r + \alpha\Delta) = 0, \tag{10c}$$

$$\text{either} \quad \kappa = 0, c_r \geq \alpha\Delta, \quad \text{or} \quad \kappa \in (0,1), c_r = \alpha\Delta \quad \text{or} \quad \kappa = 1, c_r \leq \alpha\Delta. \tag{10d}$$

In objective function (10a), the firm's total expected profit comprises the expected profit from base customers, $P\Lambda\beta$, and the expected profit from referred customers, $(P-\Delta)\frac{\Lambda\beta\kappa\alpha}{1-\kappa\alpha}$, where $\frac{\Lambda\beta\kappa\alpha}{1-\kappa\alpha}$ is the throughput of referred customers, and $P-\Delta$ is the profit from each referred customer because for each referred customer who joins, the firm pays the referring customer reward Δ. Equations (10b) and (10c) are participation constraints for base and referred customers, respectively. Conditions (10d) are incentive constraints for referrals. Three distinct cases are possible: either no customers refer ($\kappa=0, c_r>\alpha\Delta$), or all customers refer ($\kappa=1, c_r<\alpha\Delta$), or customers are indifferent to referral ($c_r=\alpha\Delta$) and thus randomize, i.e. $\kappa\in(0,1)$.

We simplify problem (10a)-(10d) to (11a)-(11c) by recognizing the profit from each customer is the effective net price $P^e \triangleq P-\kappa\alpha\Delta$, because $\kappa\alpha\Delta$ is the expected amount by which the firm compensates each customer for referrals.

$$\max_{P^e\geq 0;\ (\alpha,\beta,\kappa)\in[0,1]^3} \quad P^e\frac{\Lambda\beta}{1-\kappa\alpha} \tag{11a}$$

$$\text{s.t.} \quad \bar{V}(1-\beta)-P^e-cW^B(\alpha,\beta,\kappa)-\kappa c_r=0, \tag{11b}$$

$$\bar{V}(1-\alpha)-P^e-cW^R(\alpha,\beta,\kappa)-\kappa c_r=0. \tag{11c}$$

If the optimal solution to (11a)-(11c) requires a mixed referral strategy $\kappa\in(0,1)$, then admission price P and referral reward Δ would be uniquely determined from the optimal solution by letting $\Delta=c_r/\alpha$, and $P=P^e+\kappa\alpha\Delta$. However, if the optimal solution has either $\kappa=0$ or $\kappa=1$, then P and Δ would not be uniquely determined. If $\kappa=0$ (no-referrals), the firm can conveniently set $\Delta=0$ and $P=P^e$. If $\kappa=1$, any $\Delta\geq c_r/\alpha, P=P^e+\kappa\alpha\Delta$ would do. The idea is that charging a high price first and offering a large referral reward later would exert no effect on the profit as long as the effective net price is the same (see also Lobel et al. 2016). We call the referral reward program with optimized P and Δ the optimal referral reward program.

Proposition 7. *The optimal referral reward program generates referrals when the base market size Λ is small enough and $\bar{V}$ is high enough.*

As in the referral priority program, high service valuation $\bar{V}$ would induce referrals also in the referral reward program. However, contrary to the referral priority program (Proposition 6), which would fail to generate referrals when the base market size is too small, the optimal referral reward program actually incentivizes referrals under this scenario. The reason is that when the base market size is large, the conversion rate of referrals would be low, which makes compensating referrals too costly. Thus, the firm would not use the referral priority program and revert to the FIFO monopoly price. By contrast, a small base market size contributes to a high conversion rate, and, therefore, compensating referrals becomes cost-effective.

6.3. Numerical Comparison

We conduct a numerical study to compare the profit performance of the two referral programs. Table 1 tabulates the two programs' percentage change in profit relative to the non-referral FIFO benchmark for different base market size Λ and different maximum service valuation $\bar{V}$, fixing c, μ, c_r. For example, for $\Lambda = 0.1,\ \bar{V} = 12.5$, the referral priority program increases the profit of the non-referral FIFO benchmark by 5.49%, whereas the referral reward program increases this benchmark profit by 51.93%; in this case, the referral reward program is more profitable than the referral priority program. Table 1 can be partitioned into three regimes in terms of profit comparison. In Regime 1, both programs achieve the same profit equal to the non-referral FIFO benchmark (the lower-left cells, demarcated by the dashed lines). In Regime 2, the referral priority program outperforms the referral reward program (cells in the middle right, demarcated by the solid lines). In Regime 3, the referral priority program earns a lower profit than the referral reward program (upper-left and lower-right cells). We obtain the following observations from Table 1.

Table 1 Percentage change in profit (%) of the referral priority program (first) and the referral reward program (second) relative to the non-referral FIFO benchmark.

Λ	$\bar{V}=5$	$\bar{V}=7.5$	$\bar{V}=10$	$\bar{V}=12.5$	$\bar{V}=15$	$\bar{V}=17.5$	$\bar{V}=20$
0.1	(0, 6.47)	(0, 26.81)	(0, 41.02)	(5.49, 51.93)	(33.38, 60.76)	(51.82, 68.16)	(64.87, 74.50)
0.3	(0, 2.25)	(0, 19.48)	(26.88, 31.57)	(40.56, 40.86)	(48.42, 48.39)	(54.76, 54.71)	(60.19, 60.12)
0.5	(0, 0.27)	(4.97, 13.41)	(23.27, 23.59)	(31.43, 31.40)	(37.8, 37.74)	(43.14, 43.05)	(47.72, 47.62)
0.7	(0, 0)	(3.60, 8.58)	(17.09, 17.09)	(23.66, 23.61)	(28.97, 28.89)	(33.43, 33.33)	(37.25, 37.13)
0.9	(0, 0)	(0.78, 4.86)	(11.97, 11.95)	(17.43, 17.36)	(21.83, 21.74)	(25.52, 25.41)	(28.69, 28.57)
1.1	(0, 0)	(0, 2.12)	(8.02, 7.99)	(12.55, 12.48)	(16.19, 16.10)	(19.23, 19.13)	(21.84, 21.73)
1.3	(0, 0)	(0, 0.61)	(5.03, 4.99)	(8.81, 8.73)	(11.82, 11.73)	(14.33, 14.23)	(16.48, 16.37)
1.5	(0, 0)	(0, 0.03)	(2.79, 2.75)	(5.97, 5.90)	(8.49, 8.40)	(10.58, 10.48)	(12.35, 12.25)
1.7	(0, 0)	(0, 0)	(1.06, 1.20)	(3.84, 3.77)	(5.97, 5.89)	(7.72, 7.63)	(9.20, 9.11)
1.9	(0, 0)	(0, 0)	(0, 0.36)	(2.24, 2.17)	(4.07, 3.99)	(5.56, 5.47)	(6.81, 6.72)
2.1	(0, 0)	(0, 0)	(0, 0.02)	(1.03, 1.04)	(2.62, 2.55)	(3.91, 3.83)	(4.99, 4.91)
2.3	(0, 0)	(0, 0)	(0, 0)	(0.12, 0.37)	(1.52, 1.45)	(2.65, 2.57)	(3.59, 3.51)
2.5	(0, 0)	(0, 0)	(0, 0)	(**-0.47**, 0.06)	(0.67, 0.69)	(1.68, 1.60)	(2.51, 2.44)
2.7	(0, 0)	(0, 0)	(0, 0)	(**-0.29**, 0)	(0.02, 0.24)	(0.92, 0.87)	(1.67, 1.60)
2.9	(0, 0)	(0, 0)	(0, 0)	(**-0.13**, 0)	(**-0.49**, 0.04)	(0.33, 0.39)	(1.00, 0.94)
3.1	(0, 0)	(0, 0)	(0, 0)	(**-0.03**, 0)	(**-0.89**, 0)	(**-0.14**, 0.12)	(0.47, 0.47)

$c = 1,\ \mu = 1,\ c_r = 0.2$.

Observation 1: Customers do not refer in either program when the base market size is large and the service valuation is low (Regime 1). This observation is consistent with Propositions 6 and 7. In both schemes, a large base market size deters referrals, whereas a high service valuation stimulates referrals. Note that this regime is not displayed for $\bar{V} \geq 12.5$ in the table, because we truncate the base market size at 3.1 (this regime would appear under larger base market sizes for $\bar{V} \geq 12.5$).

Observation 2: When the base market size is intermediately small and the service valuation is relatively high (Regime 2), the referral priority program is favored over the referral reward program. In this regime, both referral programs generate referrals and achieve a higher profit than the non-referral FIFO benchmark. However, in this case, the referral priority program is more efficient in generating referrals, because it relies on customers' incentive to gain priority and does not require monetary compensation on the part of the firm. With a higher service valuation, the referral priority program outperforms the referral reward program for a wider range of base market sizes, which is reminiscent of the shape of Region III in Figure 2. A high service valuation is conducive to referrals in both programs, but the base market size is intermediately small, and referrals are more efficient in the referral priority program. Note that the magnitude of the profit difference between the two programs is relatively small. We numerically find that in this regime, all joining customers refer in both programs, making these two systems somewhat comparable.

Observation 3: The referral reward program is superior to the referral priority program either when the base market size is small or intermediately large (Regime 3). When the base market size is small, as we establish in Propositions 6 and 7, the two programs would beget opposite referral outcomes: the referral reward program encourages referrals, whereas the referral priority program fails to generate referrals. Hence, the referral reward program is more profitable. On the other hand, when the base market size is intermediately large, referrals in the referral priority program backfire in improving throughput relative to FIFO, keeping the price fixed (as our analysis in §5 demonstrates), and sometimes optimally adjusting the price may not fully counteract this adverse effect, as illustrated in those cells with boldfaced, negative profit changes (the optimal referral priority program achieves a lower profit even than the non-referral FIFO benchmark in those cases, reminiscent of the shape of Region II in Figure 2). We should note that even when the referral priority program generates referrals and improves profit over FIFO, it may still be dominated by the referral reward program. This dominance typically occurs when the base market size is small or the base market size is intermediately large. In the former case, the firm must charge a low enough price to create congestion and an incentive for priority, which makes referrals less efficient. In the latter case, the firm must use the price lever to combat the adverse effect of referrals (either a lower price to improve the conversion rate, or a higher price to follow a margin strategy), which again makes the referral priority program inferior to the referral reward program.

In Appendix B, we report our numerical findings of the firm's price adjustment and the resulting throughput changes relative to FIFO when introducing these two referral programs. Interestingly, we find that in both programs, the firm may either increase or decrease the price (the effective net price in the case of the referral reward program); in both programs, sometimes the price is increased by so much that the throughput is lower than FIFO (when the total profit is improved). This observation indicates a generic, subtle aspect of running a referral program in queueing settings: the referral program may not always serve to acquire more customers, but rather, may create an opportunity for the firm to actually raise prices high enough that fewer customers would join.

7. Concluding Remarks

As an emerging business practice, the referral priority program has been quickly adopted by a growing number of technology companies that need to waitlist customers. In such a referral priority program, customers on a waitlist can gain priority access if they successfully invite a friend to join the waitlist. This is an appealing value proposition because the firm may attract more customers without providing any monetary reward (as in the classical referral reward program). The operational characteristics of waitlists and the interdependencies of customers' referral incentives are the key underlying drivers of the referral priority program. As such, the referral priority program's effectiveness in customer acquisition and its impact on customer welfare entail further analysis.

Our paper fills this gap. We find the referral priority program does not fit all environments, depending critically on the base market size, among other market conditions. Specifically, if the base market size is too large or too small, customers do not make referrals. If the base market size is intermediately large, customers refer, but the system throughout may actually *decline* in the presence of the referral priority program, and customer welfare would deteriorate when a lower system throughput ensues. Only when the base market size is intermediately small will the program achieve a higher throughput. However, customer welfare in this case may still decrease. We also consider the firm's optimal pricing problem in the referral priority program and compare it with the referral reward program in which the firm sets both an admission price and a referral reward. Our numerical study suggests that even when the firm optimally adjusts its admission price, the referral priority program may still be less profitable than the non-referral FIFO benchmark in some cases. On the other hand, when the base market size is intermediately small and the service valuation is high, the referral priority program tends to outperform the referral reward program (which is always at least as profitable as the non-referral FIFO benchmark).

Our model highlights the advantages/disadvantages of the referral priority program. It provides one plausible theoretical explanation as to why some companies complement their waitlists with a referral priority program (e.g, Robinhood, LBRY) and some do not (e.g., Dropbox, Mailbox). Thus,

when deciding whether to introduce the seemingly innocuous referral priority program, firms should exercise discretion and conduct careful market research to understand the underlying business environment. Moreover, we also provide guidelines to waitlist managers choosing between the referral priority program and the referral reward program. While the referral priority program may not be desirable at all times, it could, under some market conditions, earn a higher profit than the referral reward program. Given the growing availability of supporting tools like Waitlisted.co and the simplicity in crediting referrals (no financial transactions involved), the referral priority program may be more favorable both in profitability and implementability under those circumstances.

Waitlists and referral programs therein provide a rich context for various research inquiries.

Customers' heterogeneity in delay sensitivity can be incorporated in our model. We can consider a model comprised of multiple classes of customers indexed by their waiting costs, in the same vein as Mendelson and Whang (1990) and Afèche (2013). For each class of customers, we need to specify the joining and referral strategies in equilibrium for both base and referred customers in the class, $(\beta, \alpha, r^B, r^R)$. One would expect the referral priority program to exploit heterogeneous delay preferences and capture more impatient customers with the priority option. However, because the firm cannot directly control referral costs as it would prices in priority pricing, the firm may have difficulty inducing different classes of customers to self-select into different priorities. Thus, in addition to within-class demand cannibalization (base customers are lost as referred customers join), between-class demand cannibalization might arise, i.e., when customers with low waiting costs refer, too much congestion might result, crowding out impatient customers.

Future research could investigate observable queues and dynamic referrals both in terms of queue-length-dependent referrals upon arrival and making referrals while waiting in the queue. Unlike priority purchasing in Adiri and Yechiali (1974) and Hassin and Haviv (1997), the referral priority program may not eventuate in a threshold-type referral strategy (by which a customer refers if the queue length upon arrival is above a certain threshold), because a long queue would deter referred customers from joining, and thus cause existing customers not to refer. When a customer is pushed back while in the queue because of other customers' referral behavior, she may be motivated to make a referral. This dynamic behavior in priority queues is reminiscent of Afèche and Sarhangian (2015), which features rational abandonment of customers in the regular class after they are overtaken by priority customers. Here, customers could also abandon, but may resort to the additional choice of inviting a friend.

By way of abstraction, our model assumes a sequential service process, whereas in many practical applications, customers may be taken off the waitlist periodically in batches. Batch service would have nuanced implications for rational customers' referral incentives, because within a batch, the relative positions of customers do not affect their expected delays. Moreover, customers in different

priority classes may be served in the same batch, diluting the attractiveness of gaining priorities. Related are questions that touch on the optimal choice of service rate: how many customers should be served in a batch, and how often should the firm trigger a batch service. These questions prevail for any waitlist even without referral programs.

Referral frauds are a particular concern for firms trying to manage a fair waitlist. Various strategies are in place to combat fake referrals or duplicate sign-ups. For instance, Robinhood credits customers with one successful invite only if a referred friend's brokerage application is approved. Opportunistic referral behaviors artificially inflate system congestion and may either motivate more referrals from other good-faith customers by creating a stronger incentive for priority, or trigger more balking of sincere customers who would otherwise join. Future research could investigate the impact of fake referrals on customer and system behavior as well as their fairness implications.

References

Adiri, I., U. Yechiali. 1974. Optimal priority-purchasing and pricing decisions in nonmonopoly and monopoly queues. *Operations Research* **22**(5) 1051–1066.

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

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

Afèche, P., V. Sarhangian. 2015. Rational abandonment from priority queues: equilibrium strategy and pricing implications. Working paper, University of Toronto.

Alperstein, H. 1988. Note—optimal pricing policy for the service facility offering a set of priority prices. *Management Science* **34**(5) 666–671.

Balachandran, K. R. 1972. Purchasing priorities in queues. *Management Science* **18**(5) 319–326.

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

Burke, P. J. 1975. Delays in single-server queues with batch input. *Operations Research* **23**(4) 830–833.

Buttle, F. A. 1998. Word of mouth: understanding and managing referral marketing. *Journal of Strategic Marketing* **6**(3) 241–254.

Chaudhry, M. L., J. G. C. Templeton. 1983. *A First Course in Bulk Queues*. Wiley, New York.

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

Garnefeld, I., A. Eggert, S. V. Helm, S. S. Tax. 2013. Growing existing customers' revenue streams through customer referral programs. *Journal of Marketing* **77**(4) 17–32.

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

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

Godes, D., D. Mayzlin, Y. Chen, S. Das, C. Dellarocas, B. Pfeiffer, B. Libai, S. Sen, M. Shi, P. Verlegh. 2005. The firm's management of social interactions. *Marketing Letters* **16**(3) 415–428.

Hamburger, E. 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`.

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

Hassin, R., M. Haviv. 1997. Equilibrium threshold strategies: The case of queues with priorities. *Operations Research* **45**(6) 966–973.

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

Hawkes, A. G. 1965. Time-dependent solution of a priority queue with bulk arrival. *Operations Research* **13**(4) 586–595.

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

Kittsteiner, T., B. Moldovanu. 2005. Priority auctions and queue disciplines that depend on processing time. *Management Science* **51**(2) 236–248.

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

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

Libai, B., E. Biyalogorsky, E. Gerstner. 2003. Setting referral fees in affiliate marketing. *Journal of Service Research* **5**(4) 303–315.

Lobel, I, E. Sadler, L. R. Varshney. 2016. Customer referral incentives and social media. *Management Science* Forthcoming.

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

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

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

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

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

Ross, S. M. 1996. *Stochastic Processes*. 2nd ed. Wiley, New York.

Ryu, G., L. Feick. 2007. A penny for your thoughts: Referral reward programs and referral likelihood. *Journal of Marketing* **71**(1) 84–94.

Schmitt, P., B. Skiera, C. Van den Bulte. 2011. Referral programs and customer value. *Journal of Marketing* **75**(1) 46–59.

Shontell, A. 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`.

Takagi, H., Y. Takahashi. 1991. Priority queues with batch poisson arrivals. *Operations Research Letters* **10**(4) 225–232.

Takahashi, Y., H. Takagi. 1990. Structured priority queue with batch arrivals. *Journal of the Operations Research Society of Japan* **33**(3) 244–263.

Trusov, M., R. E. Bucklin, K. Pauwels. 2009. Effects of word-of-mouth versus traditional marketing: Findings from an internet social networking site. *Journal of Marketing* **73**(5) 90–102.

Verlegh, P. W. J., G. Ryu, M. A. Tuk, L. Feick. 2013. Receiver responses to rewarded referrals: the motive inferences framework. *Journal of the Academy of Marketing Science* **41**(6) 669–682.

Wirtz, J., P. Chew. 2002. The effects of incentives, deal proneness, satisfaction and ties strength on word-of-mouth behaviour. *International Journal of Service Industry Management* **13**(2) 141–162.

Wolff, Ronald W. 1982. Poisson Arrivals See Time Averages. *Operations Research* **30**(2) 223–231.

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

Yildirim, U., J. J. Hasenbein. 2010. Admission control and pricing in a queue with batch arrivals. *Operations Research Letters* **38**(5) 427–431.

Ziani, S., F. Rahmoune, M. S. Radjef. 2015. Customers' strategic behavior in batch arrivals $M^2/M/1$ queue. *European Journal of Operational Research* **247**(3) 895–903.

Appendix A: Technical Proofs

Proof of Lemma 1 Let W denote the expected delay of the entire queueing system. Let W_1 and W_2 denote the expected delay in Class 1 and Class 2, respectively. Let W_i^χ denote the expected delay for customer type $\chi \in \{B, R\}$ in Class $i \in \{1, 2\}$. The main proof technique is mean value analysis. For clarity, the proof proceeds in two steps. In Step 1, We derive the expression for W; and those for other terms in step 2.

Step 1. For an arriving customer, the expected delay W has three components, her own expected service time $1/\mu$; the expected time to serve all existing customers, Q/μ, where Q is the expected queue length (by the PASTA property); and the expected delay due to other customers in the same batch, denote by W_{batch}. Thus,

$$W = \frac{1}{\mu} + \frac{Q}{\mu} + W_{\text{batch}}.$$

We derive W_{batch} by conditioning on the batch size N.

$$W_{\text{batch}} = \sum_{k=1}^{\infty} \mathbb{E}[\overline{W}_{\text{batch}} | N = k] \mathbb{P}[\text{arriving in a batch of size } k],$$

where $\mathbb{E}[\overline{W}_{\text{batch}} | N = k] = \frac{k-1}{2\mu}$ and $\mathbb{P}[\text{arriving in a batch of size } k]$ is the size-biased distribution of $\mathbb{P}[N = k]$.

$$\mathbb{P}[\text{arriving in a batch of size } k] = \frac{k\mathbb{P}[N = k]}{\mathbb{E}[N]} = \frac{kq(1-p)^2 p^{k-2}}{1+q-p}, \quad k \geq 2.$$

We use the size biased distribution to reflect the fact that a random arriving customer is more likely to be in a large batch because a large batch contains more customers. Then, by some algebra,

$$W_{\text{batch}} = \frac{q(1-p)^2}{2\mu(1+q-p)} \sum_{k=2}^{\infty} k(k-1)p^{k-2} = \frac{q}{\mu(1-p)(1+q-p)}.$$

Now, we can derive the expression of W as a function of input parameters λ, q, p.

$$\begin{aligned} W(\lambda, q, p) &= \frac{1}{\mu} + \frac{Q}{\mu} + W_{\text{batch}} \\ &= \frac{1}{\mu} + \frac{\frac{1+q-p}{1-p}\lambda W}{\mu} + \frac{q}{\mu(1-p)(1+q-p)} \qquad \text{(by Little's Law)} \\ &= \frac{1+q-p+(1-p)(q-p)}{(1+q-p)[\mu(1-p) - \lambda(1+q-p)]}. \end{aligned}$$

Step 2. Due to preemptive priority, the expected delay in Class 1, W_1, can be derived similarly as W. The first $N-1$ customers in a batch of size N join Class 1. Thus, there is an arrival to Class 1 if the batch size to the system exceeds 1, which occurs with probability q. Therefore, the arrival rate to Class 1 is λq. Conditioned on an arrival, the batch size follows a geometric distribution with parameter p. Hence,

$$W_1 = W(\lambda q, p, p) = \frac{1}{\mu(1-p) - \lambda q}.$$

By work conservation,

$$\lambda_1 W_1 + \lambda_2 W_2 = (\lambda_1 + \lambda_2) W,$$

where λ_1 and λ_2 are the throughputs to Class 1 and Class 2, respectively. $\lambda_1 = \lambda(\mathbb{E}[N] - 1) = \lambda q/(1-p)$, $\lambda_2 = \lambda$. Thus,

$$W_2 = \frac{(\lambda_1 + \lambda_2)W - \lambda_1 W_1}{\lambda_2} = \frac{\mu(1+q-p)(1-p) - \lambda q(q-p)}{[\mu(1-p) - \lambda q][\mu(1-p) - \lambda(1+q-p)]}.$$

For a base customer (the first customer in each batch) that joins Class 1, her expected delay is her own expected service time $1/\mu$ plus the expected time to serve all existing customers in Class 1, Q_1/μ, where Q_1 is the expected queue length in Class 1.

$$\omega_1^B(\lambda,q,p) = W_1^B = \frac{1}{\mu} + \frac{Q_1}{\mu} = \frac{1}{\mu} + \frac{\lambda_1 W_1}{\mu} = \frac{1}{\mu} + \frac{\frac{\lambda q}{1-p}\frac{1}{\mu(1-p)-\lambda q}}{\mu} = \frac{\mu(1-p)^2+\lambda pq}{\mu(1-p)(\mu(1-p)-\lambda q)}.$$

For a base customer that joins Class 2, her expected delay is her own expected service time $1/\mu$, plus the expected time to serve all existing customers in the system, Q/μ, plus the extra delay she expects due to overtaking of future arriving priority customers. Since she expects to spend W_1^B in the system, the expected number of priority customers who arrive in this period is $\lambda_1 W_2^B$. Thus, this extra delay due to overtaking is $\lambda_1 W_2^B/\mu$. We have

$$W_2^B = \frac{1}{\mu} + \frac{Q}{\mu} + \frac{\lambda_1 W_2^B}{\mu}.$$

Shuffling terms yields

$$\omega_2^B(\lambda,q,p) = W_2^B = \frac{1+Q}{\mu-\lambda_1} = \frac{1+\frac{1+q-p}{1-p}\lambda\frac{1+q-p+(1-p)(q-p)}{(1+q-p)[\mu(1-p)-\lambda(1+q-p)]}}{\mu-\frac{\lambda q}{1-p}} = \frac{\mu(1-p)^2+\lambda q}{[\mu(1-p)-\lambda q][\mu(1-p)-\lambda(1+q-p)]}.$$

For a referred customer (averaging over non-first customers in a batch) that joins Class 1, we can derive her expected delay by work conversation:

$$W_1^R = \frac{\lambda_1 W_1 - \lambda_1^B W_1^B}{\lambda_1^R},$$

where $\lambda_1^B = \lambda q$ is the throughput of base customers to Class 1, $\lambda_1^R = \lambda_1 - \lambda_1^B = \lambda qp/(1-p)$ is the throughput of referred customers to Class 1. Hence, plugging $\lambda_1, W_1, \lambda_1^B, W_1^B, \lambda_1^R$ gives

$$\omega_1^R(\lambda,q,p) = W_1^R = \frac{\mu(2-p)-\lambda q}{[\mu(1-p)-\lambda q]\mu}.$$

Similarly, for a referred customer (averaging over non-first customers in a batch) that joins Class 2, we can derive her expected delay by work conversation:

$$W_2^R = \frac{(\lambda_2^R+\lambda_2^B)W_2 - \lambda_2^B W_2^B}{\lambda_2^R},$$

where $\lambda_2^B = \lambda(1-q)$ is the throughput of base customers to Class 2, $\lambda_2^R = \lambda q$ is the throughput of referred customers to Class 2. After some algebra,

$$\omega_2^R(\lambda,q,p) = W_2^R = \frac{(1-p)(\mu(2-p)-\lambda)}{[\mu(1-p)-\lambda q][\mu(1-p)-\lambda(1+q-p)]}. \quad \square$$

Proof of Corollary 1 This follows from the expressions derived in Lemma 1.

$$\omega_1^R(\lambda,q,p) - \omega_1^B(\lambda,q,p) = \frac{1}{\mu(1-p)}, \quad \omega_2^R(\lambda,q,p) - \omega_2^B(\lambda,q,p) = \frac{1}{\mu(1-p)-\lambda q}.$$

Thus, $\omega_2^R(\lambda,q,p) - \omega_2^B(\lambda,q,p) \geq \omega_1^R(\lambda,q,p) - \omega_1^B(\lambda,q,p)$ with equality if and only if $q=0$. $\square$

Now, we introduce Lemma A.1 that will be used in subsequent proofs. It formally establishes that intuition that increasing either the arrival rate λ, or the unconditional probability that a base customer brings in a friend q, or the same probability for referred customers p, will increase the expected delay for both base and referred customers in both the priority and regular class.

LEMMA A.1. *The priority queue with batch arrivals has the following comparative statics:*

$$\frac{\partial \omega_i^\chi(\lambda,q,p)}{\partial \lambda}>0,\quad \frac{\partial \omega_i^\chi(\lambda,q,p)}{\partial q}>0,\quad \frac{\partial \omega_i^\chi(\lambda,q,p)}{\partial p}>0\quad \forall i=1,2,\ \chi\in\{B,R\}.$$

Proof of Lemma A.1 The signs of the partial derivatives w.r.t. q and λ are straightforward by inspection. Hence, we only show

$$\frac{\partial \omega_i^\chi(\lambda,q,p)}{\partial p}>0\quad \forall i=1,2,\ \chi\in\{B,R\}.$$

Note that both $1/(\mu(1-p))$ and $1/(\mu(1-p)-\lambda q)$ are increasing in p. Thus, we only need to prove $\omega_1^B(\lambda,q,p)$ and $\omega_2^B(\lambda,q,p)$ are increasing in p.

$$\omega_1^B(\lambda,q,p)=\frac{\mu(1-p)^2+\lambda pq}{\mu(1-p)(\mu(1-p)-\lambda q)}=\frac{\mu(1-p)^2+\lambda pq}{\mu[\mu(1-p)^2+\lambda pq]-\mu\lambda q}=\frac{1}{\mu-\frac{\mu\lambda q}{[\mu(1-p)^2+\lambda pq]}}.$$

It suffices to prove $\mu(1-p)^2+\lambda pq$ is decreasing in p.

$$\frac{\partial}{\partial p}\left[\mu(1-p)^2+\lambda pq\right]=-2\mu(1-p)+\lambda q<-2\mu(1-p)+\mu(1-p)<0.$$

This proves $\omega_1^B(\lambda,q,p)$ is increasing in p.

To prove $\omega_2^B(\lambda,q,p)$ is increasing in p, we first introduce some relabeling. Let $x\triangleq 1-p$ and $y\triangleq \lambda q$.

$$\omega_2^B(\lambda,q,p)=\frac{\mu(1-p)^2+\lambda q}{[\mu(1-p)-\lambda q][\mu(1-p)-\lambda(1+q-p)]}=\frac{\mu x^2+y}{(\mu x-y)(\mu x-\lambda x-y)}.$$

Note that $y<\mu x-\lambda x$. It suffices to show that $\frac{\mu x^2+y}{(\mu x-y)(\mu x-\lambda x-y)}$ is decreasing in x.

$$\frac{\partial}{\partial x}\left[\frac{\mu x^2+y}{(\mu x-y)(\mu x-\lambda x-y)}\right]=y\frac{\mu x[\lambda(x+2)-2\mu(x+1)]+y[2\mu(x+1)-\lambda]}{(\mu x-y)^2(\mu x-\lambda x-y)^2}.$$

We need to show that $\mu x[\lambda(x+2)-2\mu(x+1)]+y[2\mu(x+1)-\lambda]<0$. Since $y<\mu x-\lambda x$ and $2\mu(x+1)>\lambda$,

$$\begin{aligned}&\mu x[\lambda(x+2)-2\mu(x+1)]+y[2\mu(x+1)-\lambda]\\ <&\mu x[\lambda(x+2)-2\mu(x+1)]+(\mu x-\lambda x)[2\mu(x+1)-\lambda]\\ =&-\lambda x[\mu(1+x)-\lambda]<0.\end{aligned}$$

This proves that $\omega_2^B(\lambda,q,p)$ is also increasing in p. □

Proof of Proposition 1 We start the proof by determining the cutoff value c_r^l: $c_r^l=c\alpha^*[W_2^B(\alpha^*,\beta^*)-W_1^B(\alpha^*,\beta^*)]$, where α^*,β^* solve the following simultaneous equations.

$$\bar{V}(1-\beta^*)-cW_2^B(\alpha^*,\beta^*)=0, \tag{A.1a}$$

$$\bar{V}(1-\alpha^*)-c\alpha^*[W_2^B(\alpha^*,\beta^*)-W_1^B(\alpha^*,\beta^*)]-c[\alpha^*W_1^R(\alpha^*,\beta^*)+(1-\alpha^*)W_2^R(\alpha^*,\beta^*)]=0, \tag{A.1b}$$

where $W_i^\chi(\alpha,\beta)=\omega_i^\chi(\Lambda\beta,\alpha,\alpha),\ i=1,2,\ \chi\in\{B,R\}$. By construction, when $c_r=c_r^l$, base customers are indifferent to referrals. We shall show that for any $c_r\in[0,c_r^l]$, there exists a unique $(1,1)$ equilibrium. Moreover, the equilibrium $\beta(c_r)$ and $\alpha(c_r)$ are decreasing in c_r. Note that monotonicity of $\beta(c_r)$ and $\alpha(c_r)$ immediately implies that throughput $\Lambda\beta(c_r)/(1-\alpha(c_r))$ is decreasing in c_r.

Step 1: We shall show that β and α determined by (5a) and (5c) move in the same direction (either both increasing or both decreasing) as c_r changes. Subtracting (5c) from (5a) yields

$$h(\alpha,\beta) \triangleq \bar{V}(1-\beta) - \bar{V}(1-\alpha) + c\alpha[W_1^R(\alpha,\beta) - W_1^B(\alpha,\beta)] + c(1-\alpha)[W_2^R(\alpha,\beta) - W_2^B(\alpha,\beta)] = 0.$$

After simplification,

$$h(\alpha,\beta) = \bar{V}(1-\beta) - \bar{V}(1-\alpha) + \frac{c\alpha}{\mu(1-\alpha)} + \frac{c(1-\alpha)}{\mu(1-\alpha) - \Lambda\beta\alpha} = 0. \tag{A.2}$$

We shall show $d\alpha/d\beta = -\frac{\partial h/\partial\beta}{\partial h/\partial\alpha} > 0$ (which would imply β and α move in the same direction), where

$$\frac{\partial h}{\partial \alpha} = \bar{V} + \frac{c}{\mu(1-\alpha)^2} + \frac{c\Lambda\beta}{[\mu(1-\alpha) - \Lambda\beta\alpha]^2} > 0.$$

$$\frac{\partial h}{\partial \beta} = -\bar{V} + \frac{c(1-\alpha)\Lambda\alpha}{[\mu(1-\alpha) - \Lambda\beta\alpha]^2}. \tag{A.3}$$

It suffices to show $\frac{\partial h}{\partial \beta} < 0$. Since

$$\bar{V}(1-\beta) - c[\alpha W_1^B(\alpha,\beta) + (1-\alpha)W_2^B(\alpha,\beta)] = c_r \geq 0,$$

it follows that

$$\bar{V} - c(1-\alpha)W_2^B(\alpha,\beta) > 0, \tag{A.4}$$

where

$$W_2^B(\alpha,\beta) = \frac{\mu(1-\alpha)^2 + \Lambda\beta\alpha}{[\mu(1-\alpha) - \Lambda\beta\alpha][\mu(1-\alpha) - \Lambda\beta]}. \tag{A.5}$$

First, in (A.5), the denominator of the RHS satisfies $[\mu(1-\alpha) - \Lambda\beta\alpha][\mu(1-\alpha) - \Lambda\beta] < [\mu(1-\alpha) - \Lambda\beta\alpha]^2$. Second, we examine a term related to the numerator of the RHS of (A.5).

$$\mu(1-\alpha)^2 + \Lambda\beta\alpha - \Lambda\alpha \underset{\text{since } \beta\geq\alpha}{\geq} \mu(1-\alpha)^2 + \Lambda\alpha\alpha - \Lambda\alpha = (1-\alpha)[\mu(1-\alpha) - \Lambda\alpha] \underset{\text{since } \beta\geq\alpha}{\geq} (1-\alpha)[\mu(1-\alpha) - \Lambda\beta] > 0.$$

The last inequality holds because $\mu(1-\alpha) - \Lambda\beta > 0$, which is the stability condition from the expression of $W_2^B(\cdot)$. Therefore,

$$W_2^B(\alpha,\beta) = \frac{\mu(1-\alpha)^2 + \Lambda\beta\alpha - \Lambda\alpha + \Lambda\alpha}{[\mu(1-\alpha) - \Lambda\beta\alpha][\mu(1-\alpha) - \Lambda\beta]} > \frac{\Lambda\alpha}{[\mu(1-\alpha) - \Lambda\beta\alpha][\mu(1-\alpha) - \Lambda\beta]} > \frac{\Lambda\alpha}{[\mu(1-\alpha) - \Lambda\beta\alpha]^2}.$$

Combining this inequality with (A.4) gives

$$-\bar{V} + \frac{c(1-\alpha)\Lambda\alpha}{[\mu(1-\alpha) - \Lambda\beta\alpha]^2} < 0.$$

Recall from (A.3) that this is the expression for $\frac{\partial h}{\partial \beta}$. Hence, $\frac{\partial h}{\partial \beta} < 0$. This completes the proof in Step 1.

Step 2: We shall show that β and α are uniquely determined by (5a) and (5c) are both decreasing in c_r.

First, we shall show that $\alpha W_1^B(\alpha,\beta) + (1-\alpha)W_2^B(\alpha,\beta)$ is increasing in α and β. For β, it is immediate following Lemma A.1. For α, it is equivalent to showing

$$p\frac{\mu(1-p)^2 + \lambda p^2}{\mu(1-p)(\mu(1-p) - \lambda p)} + (1-p)\frac{\mu(1-p)^2 + \lambda p}{[\mu(1-p) - \lambda p][\mu(1-p) - \lambda]}$$

is increasing in p. Without loss of generality, let $\mu = 1$. Thus, $\lambda < 1-p$. Its derivative w.r.t. p is

$$\frac{\lambda\left(\lambda^3(p-2)p^3 + \lambda^2 p^2\left(2p^3 - 5p^2 + 3\right) + \lambda\left(p^4 - 5p^2 - 1\right)(p-1)^2 + \left(p^3 - 3p^2 + 2p - 2\right)(p-1)^3\right)}{(p-1)^2(\lambda + p - 1)^2(\lambda p + p - 1)^2}.$$

We shall show that for $\lambda \in (0, 1-p)$, the numerator of the above term is positive, i.e.,

$$\zeta(\lambda) \triangleq \lambda^3(p-2)p^3 + \lambda^2 p^2\left(2p^3-5p^2+3\right) + \lambda\left(p^4-5p^2-1\right)(p-1)^2 + \left(p^3-3p^2+2p-2\right)(p-1)^3 > 0.$$

Since $\zeta(1-p) = (1-p)^5 > 0$. It suffices to show that $\zeta'(\lambda) < 0$.

$$\zeta'(\lambda) = 3\lambda^2(-2+p)p^3 + 2\lambda p^2(3-5p^2+2p^3) - (1-p)^2(1+5p^2-p^4).$$

Since $\zeta'(0) = -(1-p)^2(1+5p^2-p^4) < 0$ and $\zeta(1-p) = -(1-p)^3(1+p) < 0$, it suffices to prove $\zeta'(\lambda)$ is monotone.

$$\zeta''(\lambda) = 6\lambda(-2+p)p^3 + 2p^2(3-5p^2+2p^3).$$

This is a decreasing linear function in λ. Since $\zeta''(1-p) = 2p^2(1-p)(3-3p+p^2) > 0$. $\zeta''(\lambda) > 0$ for $\lambda \in (0, 1-p)$. Hence, $\zeta'(\lambda)$ is monotonically increasing. This completes the proof for the claim that $\alpha W_1^B(\alpha,\beta) + (1-\alpha)W_2^B(\alpha,\beta)$ is increasing in α and β.

Express α as a function of β determined by $h(\alpha,\beta) = 0$, denoted by $\alpha(\beta)$. From Step 1, $\alpha(\beta)$ is an increasing function. Plugging $\alpha(\beta)$ into (5a) gives an equation solely in terms of β.

$$U(\beta) \triangleq \bar{V}(1-\beta) - c[\alpha W_1^B(\alpha(\beta),\beta) + (1-\alpha(\beta))W_2^B(\alpha(\beta),\beta)] - c_r = 0. \tag{A.6}$$

By Step 1 and the second step of Step 2, $\alpha W_1^B(\alpha(\beta),\beta) + (1-\alpha(\beta))W_2^B(\alpha(\beta),\beta)$ is increasing in β. Hence, the LHS of (A.6) is decreasing in β. Thus, β can be uniquely determined and is decreasing in c_r. Since α and β move in the same direction by Step 1, α is also unique and decreasing in c_r.

Step 3: We shall show that $W_2^B(\alpha,\beta) - W_1^B(\alpha,\beta)$ is increasing in α and β. This would imply (since $\beta(c_r)$ and $\alpha(c_r)$ is decreasing in c_r by Step 2) that for any $c_r < c_r^l$, (β,α) determined by (5a) and (5c) would satisfy $c_r < c\alpha[W_2^B(\alpha,\beta) - W_1^B(\alpha,\beta)]$, and thus $(1,1)$ would indeed be an equilibrium.

Showing that $W_2^B(\alpha,\beta) - W_1^B(\alpha,\beta)$ is increasing in α and β is equivalent to showing $\omega_2^B(\lambda,p,p) - \omega_1^B(\lambda,p,p)$ is increasing in λ and p.

$$\omega_2^B(\lambda,p,p) - \omega_1^B(\lambda,p,p) = \frac{\lambda(\lambda p^2 + \mu(1-p)^2(1+p))}{\mu(1-p)(\mu(1-p)-\lambda)(\mu(1-p)-\lambda p)}.$$

It is immediate by inspection that $\omega_2^B(\lambda,p,p) - \omega_1^B(\lambda,p,p)$ is increasing in λ.

$$\frac{\partial[\omega_2^B(\lambda,p,p) - \omega_1^B(\lambda,p,p)]}{\partial p} = \frac{\lambda\left(\lambda^3p^2 + \lambda^2\mu\left(2p^3-p^2-1\right) + \lambda\mu^2 p\left(p^3-3p+2\right) + 2\mu^3(p-1)^4\right)}{\mu(p-1)^2(\lambda+\mu(p-1))^2(\lambda p+\mu(p-1))^2}.$$

Note that $\lambda < \mu(1-p)$. It is equivalent to showing

$$g(\lambda) \triangleq \lambda^3p^2 + \lambda^2\mu\left(2p^3-p^2-1\right) + \lambda\mu^2 p\left(p^3-3p+2\right) + 2\mu^3(p-1)^4 > 0$$

for $\lambda \in (0, \mu(1-p))$. First, $g(0) = 2\mu^3(p-1)^4 > 0$ and $g(\mu(1-p)) = \mu^3(p-1)^4 > 0$. We shall show that for $\lambda \in (0,\mu(1-p))$, $g(\lambda)$ attains the minimum at $\lambda = \mu(1-p)$.

$$g'(\lambda) = 3\lambda^2p^2 + 2\lambda\mu\left(2p^3-p^2-1\right) + \mu^2p\left(p^3-3p+2\right).$$

$g'(\lambda)$ is a convex quadratic function of λ. Since $p^3-3p+2 > 0$ for $p \in (0,1)$, $g'(0) > 0$. Hence, to show $g(\lambda)$ attains the minimum point at $\lambda = \mu(1-p)$, it suffices to show $g'(\mu(1-p)) < 0$ (which implies $g'(\lambda)$ cannot cross zero twice for $\lambda \in (0,\mu(1-p))$). This is true since $g'(\mu(1-p)) = -2\mu^2(1-p)^2 < 0$. Hence, $g(\lambda)$ is increasing first and then decreasing for $\lambda \in (0,\mu(1-p))$. $g(\lambda)$ attains the minimum at $\lambda = \mu(1-p)$. Since $g(\mu(1-p)) = \mu^3(p-1)^4 > 0$, we conclude $g(\lambda) > 0$. This shows that $W_2^B(\alpha,\beta) - W_1^B(\alpha,\beta)$ is increasing in α and β.

Step 4: To guarantee that there exists $c_r^l > 0$, we specify conditions such that there exists a solution (α^*, β^*) to the system of equations (A.1a) and (A.1b). This is equivalent to a system composed of (A.1a) and $h(\alpha^*, \beta^*) = 0$. As in Step 2, plugging $\alpha^* = \alpha(\beta^*)$ from $h(\alpha^*, \beta^*) = 0$ into (A.1a) would give

$$\bar{V}(1-\beta^*) - cW_2^B(\alpha(\beta^*), \beta^*) = 0.$$

The LHS is decreasing in β^* because $\alpha(\beta^*)$ is an increasing function and $W_2^B(\alpha, \beta)$ is increasing in both α and β by Lemma A.1. If β^* is large enough, the LHS will be negative. Hence, to guarantee a (unique) solution of β^* (and thus α^*), we need $U(\beta^*_{\min}) > 0$ where $\beta^*_{\min}$ is the minimum possible value for β^*. Since $\alpha^* < \beta^*$ (since $W_i^R > W_i^B$, $i = 1, 2$) and α^* is increasing in β^*. The minimum value for β^* is obtained when $\alpha^* = 0$. Therefore, we need the following conditions:

$$\bar{V}(1-\beta^*_{\min}) - cW_2^B(0, \beta^*_{\min}) > 0 \quad \text{where} \quad h(0, \beta^*_{\min}) = 0. \tag{A.7}$$

From $h(0, \beta^*_{\min}) = 0$, we get $\beta^*_{\min} = c/(\mu\bar{V})$. Plugging this into $\bar{V}(1-\beta^*_{\min}) - cW_2^B(0, \beta^*_{\min}) > 0$ gives an upper bound on Λ, which is the expression for $\bar{\Lambda}$. □

Proof of Theorem 1 We first show that if $\Lambda \geq \bar{\Lambda}$, the only equilibrium possible is the strict non-referral equilibrium $(r^B, r^R) = (0, 0)$. From (A.7) in the proof of Proposition 1, we know that if $\Lambda \geq \bar{\Lambda}$, $\bar{V}(1-\beta^*_{\min}) - cW_2^B(0, \beta^*_{\min}) \leq 0$, where $\beta^*_{\min} = c/(\mu\bar{V})$. Note that $W_2^B(0, \beta^*_{\min}) = 1/(\mu - \Lambda\beta^*_{\min})$. Since β^F satisfies $\bar{V}(1-\beta^F) - c/(\mu - \Lambda\beta^F) = 0$, we have $\beta^F \leq \beta^*_{\min} = c/(\mu\bar{V})$ (because of the monotonicity of $\bar{V}(1-x) - c/(\mu - \Lambda x)$ in x). This future implies that

$$\frac{c}{\mu - \Lambda\beta^F} \geq \bar{V} - c/\mu.$$

From condition (4c), it follows that $\alpha = 0$. Therefore, the strict non-referral equilibrium $(r^B, r^R) = (0, 0)$ is supported. Moreover, we have already shown in Proposition 1 that if $\Lambda > \bar{\Lambda}$, the all-referral equilibrium cannot be supported. The other two types of mixed-strategy equilibria cannot be supported following the same argument. In this case $(\Lambda \geq \bar{\Lambda})$, $c_r^l = c_r^m = c_r^h = 0$.

Next, we investigate the case $\Lambda < \bar{\Lambda}$. We have already shown in Proposition 1 how to determine c_r^l and that for $c_r \in [0, c_r^l]$, there exists a unique all-referral equilibrium. We now turn to the specification of c_r^h and c_r^m.

First, define $c_r^h = c\alpha(1/(\mu - \Lambda\beta) - 1/\mu)$ where (α, β) solve the following equations

$$\begin{aligned} \bar{V}(1-\beta) - c/(\mu - \Lambda\beta) &= 0, \\ \bar{V}(1-\alpha) - c\left[\frac{1}{\mu - \Lambda\beta} + \frac{1}{\mu}\right] &= 0. \end{aligned}$$

Note that this system of equations always has a unique solution when $\Lambda < \bar{\Lambda}$. Comparing the definition of c_r^h with equilibrium conditions (4a)-(4c) immediately shows that the strict non-referral equilibrium can only be supported if $c_r \geq c_r^h$.

Second, define $c_r^m = c\alpha(1/(\mu - \Lambda\beta) - 1/\mu)$ where (α, β) solve the following equations

$$\bar{V}(1-\beta) - c/(\mu - \Lambda\beta) = 0, \tag{A.8}$$

$$\bar{V}(1-\alpha) - c\left[\frac{1}{\mu-\Lambda\beta} + \frac{1}{\mu(1-\alpha)}\right] = 0. \tag{A.9}$$

Again, this system of equations always has has a unique solution when $\Lambda < \bar{\Lambda}$. Inspecting equilibrium conditions for the weak non-referral equilibrium yields that increasing c_r decreases the randomization probability κ. Specifically, when $c_r = c_r^m$, $\kappa = 1$; when $c_r = c_r^h$, $\kappa = 0$. Thus, $c_r^m < c_r^h$. and the weak non-referral equilibrium can only be supported if $c_r \in [c_r^m, c_r^h]$. Note that, in general, we do not know which of c_r^l and c_r^m is bigger. If $c_r^l \geq c_r^m$, we would already have existence of equilibria. Otherwise, we would need to resort to the partial-referral equilibrium $(\kappa, 1)$ for $c_r \in [c_r^l, c_r^m]$.

Finally, to complete the proof of existence, we need to show that for $c_r \in [\min\{c_r^l, c_r^m\}, \max\{c_r^l, c_r^m\}]$, there always exists a partial-referral equilibrium $(\kappa, 1)$. From equilibrium conditions (7a)-(7c), we can alternatively view c_r as a function of κ. At $\kappa = 0$, $c_r = c_r^m$; at $\kappa = 1$, $c_r = c_r^l$. Thus, by continuity, for any $c_r \in [\min\{c_r^l, c_r^m\}, \max\{c_r^l, c_r^m\}]$, there exists a corresponding $\kappa \in [0, 1]$. Note that, it is possible that at some $\kappa \in (0, 1)$, the resulting c_r is outside the interval $\in [\min\{c_r^l, c_r^m\}, \max\{c_r^l, c_r^m\}]$, but at least existence of the partial-referral equilibrium is guaranteed within the interval. □

Proof of Corollary 2 This follows from the argument in Theorem 1. □

Proof of Corollary 3 This follows immediately by inspecting the equilibrium conditions for four forms of referral strategies, recognizing that $W_i^R > W_i^B,\ i = 1, 2$ in all four forms of equilibria. □

Proof of Proposition 2 From Theorem 1, when $c_r < c_r^m$, neither the strict or weak non-referral equilibrium can be sustained. Since equilibria always exist, there must be a referring equilibrium. To be specific, $c_r < c_r^m$ is

$$c_r < c\alpha\left(\frac{1}{\mu-\Lambda\beta} - \frac{1}{\mu}\right), \tag{A.10}$$

where (α, β) solve (A.8) and (A.9). From (A.9), $c/(\mu - \Lambda\beta) = \bar{V}(1-\alpha) - c/[\mu(1-\alpha)]$. Hence,

$$c\alpha\left(\frac{1}{\mu-\Lambda\beta} - \frac{1}{\mu}\right) = \alpha\left[\bar{V}(1-\alpha) - \frac{c}{\mu(1-\alpha)} - \frac{c}{\mu}\right] \triangleq m(\alpha).$$

The second derivative of $m(\alpha)$ is negative: $m''(\alpha) = -2\left(c + (1-\alpha)^3\mu\bar{V}\right)/[(1-\alpha)^3\mu] < 0$. Thus, $m(\alpha)$ a concave function in α. Moreover, the range for α is $(0, \bar{\alpha})$, where $\bar{\alpha}$ uniquely solves $\bar{V}(1-\bar{\alpha}) - c/[\mu(1-\bar{\alpha})] - c/\mu = 0$ (uniqueness is due to the LHS being a decreasing function of $\bar{\alpha}$).

Condition (A.10) can be rewritten as $c_r < m(\alpha)$. Let $\max_\alpha m(\alpha) = \tilde{c_r}$. Condition (A.10) will be satisfied if and only if $c_r < \tilde{c_r}$ and $\alpha \in (\alpha_1, \alpha_2)$, where α_1 and α_2 are the two roots to the equation $m(\alpha) = c_r$ (this equation has exactly two roots due to the convexity of $m(\alpha)$ and $c_r < \tilde{c_r}$). From equation (A.8), β is decreasing in Λ. This further implies $c/(\mu - \Lambda\beta)$ is increasing in Λ. Hence, from equation (A.9), α is decreasing in Λ. We can express Λ as a decreasing function of α, $\Lambda(\alpha)$. Thus, an intermediate α implies an intermediate $\Lambda \in [\tilde{\Lambda}_l, \tilde{\Lambda}_h]$, where $\tilde{\Lambda}_l = \Lambda(\alpha_2),\ \tilde{\Lambda}_h = \Lambda(\alpha_1)$.

Since $m(\alpha; \bar{V})$ is increasing in $\bar{V}$, by the envelope theorem, $\tilde{c_r} = \max_\alpha m(\alpha; \bar{V})$ is increasing in $\bar{V}$. This implies that given any c_r, there exists $\tilde{V}$ such that for $\bar{V} > \tilde{V}$, $c_r < \tilde{c_r}$.

Now we show that $\tilde{\Lambda}_l$ is decreasing and $\tilde{\Lambda}_h$ is increasing in $\bar{V}$. First, we claim that α_1 is decreasing and α_2 is increasing in $\bar{V}$. To show this, note that $m(\alpha)$ is increasing in $\bar{V}$, and therefore roots α_1 and α_2 to

equation $m(\alpha) = c_r$ are pushed to the extremes. Next, we show that $\tilde{\Lambda}_l$ is smaller and $\tilde{\Lambda}_h$ is larger with a higher $\bar{V}$. Since

$$\alpha_i \left[\bar{V}(1-\alpha_i) - \frac{c}{\mu(1-\alpha_i)} - \frac{c}{\mu} \right] = c_r, \quad i = 1, 2, \ \alpha_1 < \alpha_2$$

and α_1 is smaller whereas α_2 is larger with a higher $\bar{V}$, it follows that $\bar{V}(1-\alpha_1) - c/[\mu(1-\alpha_1)]$ is larger and $\bar{V}(1-\alpha_2) - c/[\mu(1-\alpha_2)]$ is smaller. This implies $c/(\mu - \tilde{\Lambda}_h\beta_h)$ is larger and $c/(\mu - \tilde{\Lambda}_l\beta_l)$ is smaller. Since $\bar{V}(1-\beta_l) = c/(\mu - \tilde{\Lambda}_l\beta_l)$ from (A.8), $\bar{V}(1-\beta_l)$ must be smaller. Furthermore, $\bar{V}$ is larger, which implies β_l is larger (in order for $\bar{V}(1-\beta_l)$ to be smaller). Together with $c/(\mu - \tilde{\Lambda}_l\beta_l)$ being smaller, it follows that $\tilde{\Lambda}_l$ is smaller with a higher $\bar{V}$. Finally, we turn to showing $\tilde{\Lambda}_h$ is larger with a higher $\bar{V}$. Since α_1 is smaller, it follows that

$$\frac{c}{\bar{V}(\mu - \tilde{\Lambda}_h\beta_h)} = 1 - \alpha_1 - \frac{c}{\bar{V}\mu(1-\alpha_1)}$$

is larger. From (A.8), $1 - \beta_h = c/[\bar{V}(\mu - \tilde{\Lambda}_h\beta_h)]$. Therefore, β_h is smaller. Moreover, we have shown that $c/(\mu - \tilde{\Lambda}_h\beta_h)$ is larger. This implies that $\tilde{\Lambda}_h$ is larger. □

Proof of Proposition 3 We show this result holds for both the partial-referral equilibrium (in Step 1) and the all-referral equilibrium (in Step 2), respectively.

Step 1 : For the partial-referral equilibrium, refer to the equilibrium conditions in (7a)-(7b). Since base customers are indifferent to referrals, (7a) is equivalent to

$$\bar{V}(1-\beta^R) - cW_2^B(\alpha, \beta^R, \kappa) = 0, \tag{A.11}$$

where we use notation β^R to emphasize this is for a referral equilibrium. $W_2^B(\alpha, \beta^R, \kappa) = \omega_2^B(\Lambda\beta^R, \kappa\alpha, \alpha)$. By the monotonicity property in Lemma A.1, we have $W_2^B(\alpha, \beta^R, \kappa) \geq W_2^B(0, \beta^R, 0) = 1/(\mu - \Lambda\beta^R)$ with equality at $\kappa = 0$ (which is the boundary case equivalent to non-referrals). Thus, excluding $\kappa = 0$, we have

$$\bar{V}(1-\beta^R) - c\frac{1}{\mu - \Lambda\beta^R} > 0. \tag{A.12}$$

By comparison, $\bar{V}(1-\beta^F) - c/(\mu - \Lambda\beta^F) = 0$. Hence, $\bar{V}(1-\beta^R) - c\frac{1}{\mu-\Lambda\beta^R} = \bar{V}(1-\beta^F) - c/(\mu - \Lambda\beta^F)$. Recognizing that $\bar{V}(1-x) - c/(\mu - \Lambda x)$ is decreasing in x gives implies $\beta^R < \beta^F$.

Step 2 : Consider the all-referral equilibrium. By Proposition 1, any β^R is less β^R under $c_r = 0$. Thus, it suffices to show $\beta^R < \beta^F$ when $c_r = 0$. From equilibrium condition (5a),

$$\bar{V}(1-\beta^R) - c[\alpha W_1^B(\alpha, \beta^R) + (1-\alpha)W_2^B(\alpha, \beta^R)] = 0.$$

From the proof of Proposition 1 (Step 2), $\alpha W_1^B(\alpha, \beta^R) + (1-\alpha)W_2^B(\alpha, \beta^R)$ is increasing in α. Hence,

$$\alpha W_1^B(\alpha, \beta^R) + (1-\alpha)W_2^B(\alpha, \beta^R) > W_2^B(0, \beta^R) = \frac{1}{\mu - \Lambda\beta^R}.$$

Again, we arrive at inequality (A.12). The rest of the proof is similar to the partial-referral case. □

Next, we introduce Lemma A.2 that will be used in the proof of Theorem 2, Theorem 3, and Proposition 4.

LEMMA A.2. *(i) For a partial-referral equilibrium ($\beta^R, \alpha, r^B = \kappa, r^R = 1$), a necessary and sufficient condition for its throughput to be lower than that under FIFO ($\lambda^R < \lambda^F$) is*

$$\beta^R < \frac{(1-\alpha)[2-\alpha-(1-\alpha)\mu/\Lambda]}{(2-\alpha)(1-\alpha)+\kappa\alpha}. \tag{A.13}$$

(ii) An all-referral equilibrium's throughput is lower than that under FIFO (λ^F) only if

$$\beta^R < \frac{(1-\alpha)[2-\alpha-(1-\alpha)\mu/\Lambda]}{(2-\alpha)(1-\alpha)+\alpha}, \tag{A.14}$$

where β^R, α is the equilibrium joining probabilities under c_r^l.

Proof of Lemma A.2 *Part (i)*: Under FIFO, $\lambda^F = \Lambda\beta^F$, where β^F solves $\bar{V}(1-\beta^F) - c/(\mu - \Lambda\beta^F) = 0$. Under the partial-referral equilibrium, the throughput $\lambda^R = \Lambda\beta^R[1+\kappa\alpha/(1-\alpha)]$, where (β, α^R) solves (A.11). In order for $\lambda^R < \lambda^F$, we need to have $\beta^R/(1-\alpha) < \beta^F$. This holds if and only if

$$\bar{V}\left(1-\beta^R\left[1+\kappa\frac{\alpha}{1-\alpha}\right]\right) - \frac{c}{\mu-\Lambda\beta^R[1+\kappa\alpha/(1-\alpha)]} > \bar{V}\left(1-\beta^F\right) - \frac{c}{\mu-\Lambda\beta^F} = 0.$$

Since $\bar{V}(1-\beta^R) - cW_2^B(\alpha, \beta^R, \kappa) = 0$ by (A.11), this implies

$$\bar{V}\left(1-\beta^R\left[1+\kappa\frac{\alpha}{1-\alpha}\right]\right) - \frac{c}{\mu-\Lambda\beta^R[1+\kappa\alpha/(1-\alpha)]} > \bar{V}\left(1-\beta^R\right) - cW_2^B\left(\alpha, \beta^R, \kappa\right). \tag{A.15}$$

Plugging in $W_2^B(\alpha, \beta^R, \kappa) = \frac{\mu(1-\alpha)^2+\Lambda\beta^R\kappa\alpha}{[\mu(1-\alpha)-\Lambda\beta^R\kappa\alpha][\mu(1-\alpha)-\Lambda\beta^R(1+\kappa\alpha-\alpha)]}$ and collecting terms yields

$$\frac{c}{\mu(1-\alpha)-\Lambda\beta^R(1+\kappa\alpha-\alpha)}\left[\frac{\mu(1-\alpha)^2+\Lambda\beta^R\kappa\alpha}{\mu(1-\alpha)-\Lambda\beta^R\kappa\alpha} - (1-\alpha)\right] > \frac{\bar{V}\beta^R\kappa\alpha}{1-\alpha}.$$

This simplifies to

$$\frac{c}{\mu(1-\alpha)-\Lambda\beta^R(1+\kappa\alpha-\alpha)}\left[\frac{\Lambda\beta^R\kappa\alpha(2-\alpha)}{\mu(1-\alpha)-\Lambda\beta^R(1+\kappa\alpha-\alpha)}\right] > \frac{\bar{V}\beta^R\kappa\alpha}{1-\alpha}.$$

Further algebra gives

$$\frac{c\Lambda(2-\alpha)(1-\alpha)}{[\mu(1-\alpha)-\Lambda\beta^R\kappa\alpha][\mu(1-\alpha)-\Lambda\beta^R(1+\kappa\alpha-\alpha)]} > \bar{V}.$$

Since $\bar{V}(1-\beta^R) - cW_2^B(\alpha, \beta^R, \kappa) = 0$,

$$\bar{V} = \frac{c[\mu(1-\alpha)^2+\Lambda\beta^R\kappa\alpha]}{(1-\beta^R)[\mu(1-\alpha)-\Lambda\beta^R\kappa\alpha][\mu(1-\alpha)-\Lambda\beta^R(1+\kappa\alpha-\alpha)]}, \quad \kappa \in (0,1].$$

By substitution,

$$\frac{c\Lambda(2-\alpha)(1-\alpha)}{[\mu(1-\alpha)-\Lambda\beta^R\kappa\alpha][\mu(1-\alpha)-\Lambda\beta^R(1+\kappa\alpha-\alpha)]} > \frac{c[\mu(1-\alpha)^2+\Lambda\beta^R\kappa\alpha]}{(1-\beta^R)[\mu(1-\alpha)-\Lambda\beta^R\kappa\alpha][\mu(1-\alpha)-\Lambda\beta^R(1+\kappa\alpha-\alpha)]}.$$

This simplifies to

$$(1-\beta^R) > \frac{\mu(1-\alpha)^2+\Lambda\beta^R\kappa\alpha}{\Lambda(2-\alpha)(1-\alpha)}.$$

Collecting terms gives

$$\beta^R < \frac{(1-\alpha)[2-\alpha-(1-\alpha)\mu/\Lambda]}{(2-\alpha)(1-\alpha)+\kappa\alpha}.$$

Part (ii): If an all-referral equilibrium's throughput is lower than that under FIFO, then $\lambda^R < \lambda^F$, where λ_R is the throughput is under c_r^l. This is because the throughput under c_r^l is the smallest among all-referral equilibria by Proposition 1. Also note that at c_r^l, $\bar{V}(1-\beta^R) - cW_2^B(\alpha, \beta^R) = 0$. Therefore, the result follows from the same argument in part (i) by letting $\kappa = 1$. □

Proof of Theorem 2 Combining Part (i) and (ii) in Lemma A.2, we know that any referring equilibrium would achieve a lower throughput than FIFO only if

$$\beta^R < \frac{(1-\alpha)[2-\alpha-(1-\alpha)\mu/\Lambda]}{(2-\alpha)(1-\alpha)+\kappa\alpha},$$

where $(\beta, \alpha, r^B = \kappa, r^R = 1)$ is the equilibrium. Since $\beta > \alpha$ by Corollary 3, this holds only if

$$\alpha < \frac{(1-\alpha)[2-\alpha-(1-\alpha)\mu/\Lambda]}{(2-\alpha)(1-\alpha)+\kappa\alpha}.$$

for some $\alpha \in (0,1)$. This is equivalent to $\rho(-\alpha^3+(4-\kappa)\alpha^2-5\alpha+2)-(1-\alpha)^2 > 0$, where $\rho = \Lambda/\mu$. Now we show that if $\rho \le 1/2$, then for any $\alpha \in (0,1)$ and any $\kappa \in (0,1]$, $\rho(-\alpha^3+(4-\kappa)\alpha^2-5\alpha+2)-(1-\alpha)^2 < 0$. It suffices to prove that if $\rho \le 1/2$, $\rho(-\alpha^3+4\alpha^2-5\alpha+2)-(1-\alpha)^2 < 0$, $\forall \alpha \in (0,1)$. Since $-\alpha^3+4\alpha^2-5\alpha+2 = (1-a)^2(2-a)$, we only need to prove $\rho(2-a) < 1, \forall \alpha \in (0,1)$. This is obviously true for $\rho \le 1/2$. □

Proof of Theorem 3 We prove this by finding a sufficient condition for condition (A.13) in Lemma A.2 when $\kappa = 0$. When $\kappa = 0$, condition (A.13) becomes

$$\beta^R < \frac{(1-\alpha)[2-\alpha-(1-\alpha)\mu/\Lambda]}{(2-\alpha)(1-\alpha)} = 1 - \frac{(1-\alpha)\mu/\Lambda}{(2-\alpha)}.$$

Collecting terms gives

$$\frac{1}{\mu} + \frac{1}{\mu(1-\alpha)} > \frac{1}{\Lambda(1-\beta^R)}.$$

A sufficient condition for this is

$$\frac{2}{\mu} > \frac{1}{\Lambda(1-\beta)}, \quad \text{or equivalently} \quad \bar{V}(1-\beta) > \frac{\mu\bar{V}}{2\Lambda}.$$

Since $\bar{V}(1-\beta) - c/(\mu-\Lambda\beta) = 0$ at $\kappa = 0$, it follows that

$$\frac{c}{\mu-\Lambda\beta} > \frac{\mu\bar{V}}{2\Lambda}, \quad \text{which gives} \quad \beta > \frac{\mu}{\Lambda} - \frac{2c}{\mu\bar{V}}.$$

Since $\bar{V}(1-x) - c/(\mu-\Lambda x)$ is decreasing in x and $\bar{V}(1-\beta) - c/(\mu-\Lambda\beta) = 0$, we have

$$\bar{V}\left(1 - \left(\frac{\mu}{\Lambda} - \frac{2c}{\mu\bar{V}}\right)\right) - \frac{\mu\bar{V}}{2\Lambda} > 0.$$

Simplifying this gives

$$\Lambda > \mu\frac{3\bar{V}}{2(\bar{V}+2c/\mu)}.$$

Moreover, for (β^R, α) to be sustained in equilibrium (i.e., to guarantee $\alpha > 0$), we require (see Proposition 1)

$$\mu\frac{3\bar{V}}{2(\bar{V}+2c/\mu)} < \bar{\Lambda} = \mu\frac{\bar{V}(\bar{V}-2c/\mu)}{(\bar{V}-c/\mu)c/\mu},$$

which gives $\bar{V} > 5c/2\mu$. □

Proof of Proposition 4 When $c_r = 0$, by Corollary 2, if customers refer, it is the all-referral equilibrium, which requires

$$\bar{V}(1-\beta^R) - c[\alpha W_1^B(\alpha,\beta^R) + (1-\alpha)W_2^B(\alpha,\beta^R)] = 0.$$

Following the logic of Lemma A.2, we would have a similar inequality as (A.15):

$$\bar{V}\left(1 - \frac{\beta^R}{1-\alpha}\right) - \frac{c}{\mu-\Lambda\beta^R/(1-\alpha)} > \bar{V}(1-\beta^R) - c[\alpha W_1^B(\alpha,\beta^R) + (1-\alpha)W_2^B(\alpha,\beta^R)].$$

By similar algebra as in Lemma A.2, this simplifies to

$$\frac{\beta^R\left(\alpha^4\mu(\beta^R\Lambda-\mu)+\alpha^3\left((\beta^R)^2\Lambda^2-2\beta^R\Lambda\mu+3\mu^2\right)+\alpha^2\mu(\beta^R\Lambda-4\mu)+3\alpha\mu^2-\mu^2\right)}{(\alpha-1)(\beta^R-1)\left(\alpha^3\mu(\beta^R\Lambda-\mu)+\alpha^2\left((\beta^R)^2\Lambda^2-3\beta^R\Lambda\mu+2\mu^2\right)+\alpha\mu(3\beta^R\Lambda-\mu)-\beta^R\Lambda\mu\right)}<1. \quad \text{(A.16)}$$

From Proposition 1, at $\bar{\Lambda}$, $\alpha=0$. Also note from the proof of Lemma 1 (Step 4) that $\beta^R=c/(\mu\bar{V})$ at $\Lambda=\bar{\Lambda}$. Plugging $\alpha=0,\beta^R=c/(\mu\bar{V})$ and $\bar{\Lambda}=\mu\frac{\bar{V}(\bar{V}-2c/\mu)}{(\bar{V}-c/\mu)c/\mu}$ to inequality (A.16) gives $\bar{V}>3c/\mu$. □

Proof of Proposition 5 Total customer welfare is

$$CW=\Lambda\beta\left\{\int_{\bar{V}(1-\beta)}^{\bar{V}}u^B(v,\boldsymbol{s})\frac{1}{\bar{V}}\mathrm{d}v+\frac{r^B\alpha}{1-\alpha}\int_{\bar{V}(1-\alpha)}^{\bar{V}}u^R(v,\boldsymbol{s})\frac{1}{\bar{V}}\mathrm{d}v\right\}.$$

In CW, $u^B(v,\boldsymbol{s})$ and $u^R(v,\boldsymbol{s})$ is linearly increasing in v. Since $u^B(\bar{V}(1-\beta),\boldsymbol{s})=u^R(\bar{V}(1-\alpha),\boldsymbol{s})=0$,

$$u^B(v,\boldsymbol{s})=v-\bar{V}(1-\beta),\quad u^R(v,\boldsymbol{s})=v-\bar{V}(1-\alpha).$$

Therefore,

$$\int_{\bar{V}(1-\beta)}^{\bar{V}}u^B(v,\boldsymbol{s})\frac{1}{\bar{V}}\mathrm{d}v=\frac{\bar{V}\beta^2}{2},\quad \int_{\bar{V}(1-\alpha)}^{\bar{V}}u^R(v,\boldsymbol{s})\frac{1}{\bar{V}}\mathrm{d}v=\frac{\bar{V}\alpha^2}{2}.$$

Hence, individual customer welfare is equal to

$$ICW=\frac{CW}{\Lambda\beta\left[1+\frac{r^B\alpha}{1-\alpha}\right]}=\frac{\bar{V}}{2}\left[\frac{1}{1+\frac{r^B\alpha}{1-\alpha}}\beta^2+\frac{\frac{r^B\alpha}{1-\alpha}}{1+\frac{r^B\alpha}{1-\alpha}}\alpha^2\right].$$

In a referring equilibrium (either partial or all), since $\beta^R>\alpha$ and the term in the bracket is a convex combination of $(\beta^R)^2$ and α^2, we have

$$ICW^R<\frac{\bar{V}}{2}\left(\beta^R\right)^2.$$

However, under FIFO,

$$ICW^F=\frac{\bar{V}}{2}\left(\beta^F\right)^2.$$

By Proposition 3, $\beta^R<\beta^F$, which implies $ICW^R<ICW^F$. This proves Part (i). Part (ii) immediately follows from Part (i). □

Proof of Proposition 6 We have shown in Proposition 2 that when Λ is too small or too large, customers do not refer. This is a result that holds under zero admission price, i.e., $P=0$. Making the price positive $P>0$ (which the firm would do) would only make customers less willing to join (as if customer valuation were decreased, cf. Proposition 2), decreasing the conversion rate and the need for priority. Therefore, customers would also not refer. □

Proof of Lemma 2 The proof is similar to that of Lemma 1. We prove a more general result with $\lambda=\Lambda\beta$, $q=r^B\alpha$, $p=r^R\alpha$. As in the proof of Lemma 1, let Q be the expected queue length, and W be the expected waiting time of the system, i.e., $W=\frac{1+q-p+(1-p)(q-p)}{(1+q-p)[\mu(1-p)-\lambda(1+q-p)]}$ from the proof of Lemma 1. By the PASTA property,

$$W^B=\frac{1}{\mu}+\frac{Q}{\mu}=\frac{1}{\mu}+\frac{\frac{1+q-p}{1-p}\lambda W}{\mu}=\frac{1}{\mu}+\frac{\frac{1+q-p}{1-p}\lambda\frac{1+q-p+(1-p)(q-p)}{(1+q-p)[\mu(1-p)-\lambda(1+q-p)]}}{\mu}=\frac{\mu(1-p)^2+\lambda q}{\mu\,(1-p)\,[\mu(1-p)-\lambda(1+q-p)]}.$$

Let $\lambda^B = \lambda$ and $\lambda^R = \frac{\lambda q}{1-p}$ be the throughput for base customers and referred customers, respectively. By work conservation,

$$\lambda^B W^B + \lambda^R W^R = (\lambda^B + \lambda^R) W.$$

This gives

$$W^R = \frac{\mu(2-p) - \lambda}{\mu[\mu(1-p) - \lambda(1+q-p)]}, \text{ or } W^R = W^B + \frac{1}{\mu(1-p)}. \quad \square$$

Proof of Proposition 7 By substitution, we rewrite the firm's optimization problem (11a)-(11c):

$$\max_{\alpha,\kappa} \quad \Pi\Lambda = \left[\overline{V}(1-\beta(\alpha,\kappa)) - cW^B(\alpha,\beta(\alpha,\kappa),\kappa) - \kappa c_r\right] \frac{\Lambda\beta(\alpha,\kappa)}{1-\kappa\alpha},$$

where $\beta(\alpha,\kappa) = \alpha + \frac{c}{\overline{V}\mu(1-\kappa\alpha)}$, which is obtained from subtracting (11c) from (11b).

$$\frac{\partial \Pi}{\partial \kappa} = \left[\overline{V}(1-\beta) - cW^B - \kappa c_r\right] \frac{\partial\left(\frac{\beta}{1-\kappa\alpha}\right)}{\partial \kappa} - \frac{\beta}{1-\kappa\alpha}\left[\overline{V}\frac{\partial \beta}{\partial \kappa} + c\left(\frac{\partial W^B}{\partial \kappa} + \frac{\partial W^B}{\partial \beta}\frac{\partial \beta}{\partial \kappa}\right) + c_r\right].$$

Since $\frac{\partial \beta}{\partial \kappa} = \frac{c\alpha}{\overline{V}\mu(1-\kappa\alpha)^2}$ from the expression of $\beta(\alpha,\kappa)$,

$$\frac{\partial\left(\frac{\beta}{1-\kappa\alpha}\right)}{\partial \kappa} = \frac{\frac{\partial \beta}{\partial \kappa}(1-\kappa\alpha) + \alpha\beta}{(1-\kappa\alpha)^2} = \frac{\frac{\partial \beta}{\partial \kappa}}{(1-\kappa\alpha)} + \frac{\alpha\beta}{(1-\kappa\alpha)^2} = \frac{c\alpha}{\overline{V}\mu(1-\kappa\alpha)^3} + \frac{\alpha\beta}{(1-\kappa\alpha)^2}.$$

When $\kappa = 0$,

$$\alpha = \beta - \frac{c}{\overline{V}\mu}, \; W^B = \frac{1}{\mu - \Lambda\beta}, \; \frac{\partial W^B}{\partial \kappa} = \frac{\alpha\beta\Lambda\left(2\mu - \Lambda\beta\right)}{\mu(\mu-\Lambda\beta)^2}, \; \frac{\partial W^B}{\partial \beta} = \frac{\Lambda}{\left(\mu - \Lambda\beta\right)^2}.$$

$$\begin{aligned}
\frac{\partial \Pi}{\partial \kappa}|_{\kappa=0} &= \left[\overline{V}(1-\beta) - \frac{c}{\mu - \Lambda\beta}\right]\left[\frac{c\alpha}{\overline{V}\mu} + \alpha\beta\right] - \beta\left[\frac{c\alpha}{\mu} + c\left(\frac{\partial W^B(\alpha,\beta,\kappa)}{\partial \kappa} + \frac{\partial W^B(\alpha,\beta,\kappa)}{\partial \beta}\frac{\partial \beta}{\partial \kappa}\right) + c_r\right] \\
&= \left[\overline{V}(1-\beta) - \frac{c}{\mu - \Lambda\beta}\right]\left[\frac{c\alpha}{\overline{V}\mu} + \alpha\beta\right] - \beta\left[\frac{c\alpha}{\mu} + c\left[\frac{\alpha\beta\Lambda\left(2\mu - \Lambda\beta\right)}{\mu(\mu-\Lambda\beta)^2} + \frac{\Lambda}{\left(\mu-\Lambda\beta\right)^2}\frac{c\alpha}{\overline{V}\mu}\right] + c_r\right] \\
&= \left[\overline{V}(1-\beta) - \frac{c}{\mu - \Lambda\beta}\right]\alpha\left[\beta + \frac{c}{\overline{V}\mu}\right] - \beta\frac{c}{\mu}\alpha\left[1 + \frac{\beta\Lambda\left(2\mu - \Lambda\beta\right)}{(\mu-\Lambda\beta)^2} + \frac{c\Lambda}{\overline{V}\left(\mu-\Lambda\beta\right)^2}\right] - \beta c_r \\
&= \underbrace{\alpha\left[\left[\overline{V}(1-\beta) - \frac{c}{\mu - \Lambda\beta}\right]\left(\beta + \frac{c}{\overline{V}\mu}\right) - \beta\frac{c}{\mu}\left[1 + \frac{\beta\Lambda\left(2\mu - \Lambda\beta\right)}{(\mu-\Lambda\beta)^2} + \frac{c\Lambda}{\overline{V}\left(\mu-\Lambda\beta\right)^2}\right]\right]}_{\triangleq \xi(\beta)} - \beta c_r
\end{aligned}$$

At $\kappa = 0$, β maximizes (9a)-(9b). The first-order condition yields:

$$\overline{V}(1-2\beta) - \frac{c\mu}{\left(\mu - \Lambda\beta\right)^2} = 0. \tag{A.17}$$

Note that β is decreasing in Λ. From $\alpha = \beta - \frac{c}{\overline{V}\mu}$ and $\alpha \geq 0$, we have the lower bound for β: $\beta \geq \frac{c}{\overline{V}\mu}$. Setting $\Lambda = 0$ in (A.17) give the upper bound for β: $\beta \leq \frac{1}{2} - \frac{c}{2\overline{V}\mu}$.

$$\frac{c}{\overline{V}\mu} \leq \beta \leq \frac{1}{2} - \frac{c}{2\overline{V}\mu}, \quad \bar{V} > 3c/\mu.$$

At the upper bound $\bar{\beta} = \frac{1}{2} - \frac{c}{2\overline{V}\mu} \in (\frac{1}{3}, \frac{1}{2})$ (this upper bound $\bar{\beta}$ is increasing in $\bar{V}$),

$$\bar{\alpha} = \frac{1}{2} - \frac{3c}{2\overline{V}\mu}, \quad \xi(\bar{\beta}) = \frac{(\bar{V} - c/\mu)^2}{4\bar{V}}.$$

$$\frac{\partial \Pi}{\partial \kappa}\mid_{\kappa=0;\Lambda=0}=\left(\frac{1}{2}-\frac{3c}{2\bar{V}\mu}\right)\frac{(\bar{V}-c/\mu)^2}{4\bar{V}}-\left(\frac{1}{2}-\frac{c}{2\bar{V}\mu}\right)c_r=\frac{(\bar{V}-c/\mu)^2(\bar{V}-3c/\mu)}{8\bar{V}^2}-\frac{\bar{V}-c/\mu}{2\bar{V}}c_r.$$

Or equivalently,

$$\frac{\partial \Pi}{\partial \kappa}\mid_{\kappa=0;\Lambda=0}=\frac{c}{\mu}\bar{\beta}^2\frac{3\bar{\beta}-1}{1-2\bar{\beta}}-\bar{\beta}c_r. \tag{A.18}$$

It is easy to see that $\bar{\beta}^2\frac{3\bar{\beta}-1}{1-2\bar{\beta}}$ is increasing in $\bar{\beta}\in(1/3,1/2)$. Now we show that is convex in $\bar{\beta}\in(1/3,1/2)$. Since

$$\left(\bar{\beta}^2\frac{3\bar{\beta}-1}{1-2\bar{\beta}}\right)''=2\frac{12\bar{\beta}^3-18\bar{\beta}^2+9\bar{\beta}-1}{(1-2\bar{\beta})^3},$$

It suffices to show that $12\beta^3-18\beta^2+9\beta-1>0$ for $\bar{\beta}\in(1/3,1/2)$.

$$12\bar{\beta}^3-18\bar{\beta}^2+9\bar{\beta}-1=12\bar{\beta}^3-4\bar{\beta}^2-14\bar{\beta}^2+9\bar{\beta}-1=4\bar{\beta}^2(3\bar{\beta}-1)+(1-2\bar{\beta})(7\bar{\beta}-1)>0.$$

Therefore, from (A.18), $\partial\Pi/\partial\kappa\mid_{\kappa=0;\Lambda=0}>0$ if and only if

$$\frac{c}{\mu}\bar{\beta}\frac{3\bar{\beta}-1}{1-2\bar{\beta}}>c_r. \tag{A.19}$$

Since the left-hand side of (A.19) is increasing in $\bar{\beta}$, and $\bar{\beta}$ is increasing in $\bar{V}$, we conclude that this will be satisfied when $\bar{V}$ is high enough. Also, note that $\bar{\beta}$ corresponds to $\Lambda=0$. By continuity, if (A.19) holds, there must exists $\epsilon>0$ such that for $\Lambda\in(0,\epsilon)$, $\partial\Pi/\partial\kappa\mid_{\kappa=0}>0$, which implies the optimal $\kappa^*>0$, i.e., referrals are generated. □

Appendix B: More Details on the Comparison of the Two Referral Programs

B.1. Detailed Formulation of the Optimal Referral Priority Program

Let $W_i^{\chi}(\alpha,\beta,r^B,r^R)=\omega_i^{\chi}(\Lambda\beta,r^B\alpha,r^R\alpha)$. $i=1,2,\chi\in\{B,R\}$. The conceptual model for the firm's optimal pricing problem is:

$$\begin{aligned}
\max_{P\geq 0;(\alpha,\beta,r^B,r^R)\in[0,1]^4}\quad & P\Lambda\beta\left[1+\frac{r^B\alpha}{1-\alpha}\right]\\
\text{s.t.}\quad & \bar{V}(1-\beta)-P-r^B\left(c_r-c\alpha[W_2^B(\alpha,\beta,r^B,r^R)-W_1^B(\alpha,\beta,r^B,r^R)]\right)-cW_2^B(\alpha,\beta,r^B,r^R)=0,\\
& \bar{V}(1-\alpha)-P-r^R\left(c_r-c\alpha[W_2^R(\alpha,\beta,r^B,r^R)-W_1^R(\alpha,\beta,r^B,r^R)]\right)-cW_2^R(\alpha,\beta,r^B,r^R)=0,\\
\text{either}\quad & r^B=r^R=0,\ c_r\geq c\alpha\left(\frac{1}{\mu-\Lambda\beta}-\frac{1}{\mu}\right),\\
\text{or}\quad & r^B=r^R=1,\ c_r\leq c\alpha\left[W_2^B(\alpha,\beta,1,1)-W_1^B(\alpha,\beta,1,1)\right],\\
\text{or}\quad & r^B=0,\ c_r=c\alpha\left(\frac{1}{\mu-\Lambda\beta}-\frac{1}{\mu}\right),\\
\text{or}\quad & r^R=1,\ c_r=c\alpha\left[W_2^B(\alpha,\beta,r^B,1)-W_1^B(\alpha,\beta,r^B,1)\right].
\end{aligned}$$

We operationalize this conceptual model by solving the four optimization problems below and choosing the optimal solution and value to the one that yields that the maximum objective value among the four. (An alternative approach is to introduce binary integer variables to represent the conditional statements. Given that we are already faced with nonlinear programming problems, we prefer to avoid integer variables.)

B.1.1. Referral Strategy (i): $(r^B, r^R) = (0,0)$.

$$\begin{aligned}
\max_{P,\alpha,\beta} \quad & P\Lambda\beta \\
\text{s.t.} \quad & \bar{V}(1-\beta) - \frac{c}{\mu - \Lambda\beta} - P = 0, \\
& c_r \geq c\alpha\left(\frac{1}{\mu-\Lambda\beta} - \frac{1}{\mu}\right), \\
& \text{If } \ \bar{V} - c\left[\frac{1}{\mu-\Lambda\beta} + \frac{1}{\mu}\right] - P < 0, \ \alpha = 0; \quad \text{otherwise, } \ \bar{V}(1-\alpha) - c\left[\frac{1}{\mu-\Lambda\beta} + \frac{1}{\mu}\right] - P = 0.
\end{aligned}$$

B.1.2. Referral Strategy (ii): $(r^B, r^R) = (1,1)$.

$$\begin{aligned}
\max_{P,\alpha,\beta} \quad & P\frac{\Lambda\beta}{1-\alpha} \\
\text{s.t.} \quad & \bar{V}(1-\beta) - c_r - c[\alpha W_1^B(\alpha,\beta) + (1-\alpha)W_2^B(\alpha,\beta)] - P = 0, \\
& c_r \leq c\alpha\left[W_2^B(\alpha,\beta) - W_1^B(\alpha,\beta)\right], \\
& \bar{V}(1-\alpha) - c_r - c[\alpha W_1^R(\alpha,\beta) + (1-\alpha)W_2^R(\alpha,\beta)] - P = 0,
\end{aligned}$$

where $W_i^\chi(\alpha,\beta) = \omega_i^\chi(\Lambda\beta,\alpha,\alpha), \quad i = 1,2, \ \chi \in \{B,R\}$.

B.1.3. Referral Strategy (iii): $(r^B, r^R) = (0,\kappa)$.

$$\begin{aligned}
\max_{P,\alpha,\beta,\kappa} \quad & P\Lambda\beta \\
\text{s.t.} \quad & \bar{V}(1-\beta) - c/(\mu - \Lambda\beta) - P = 0, \\
& c_r = c\alpha\left(\frac{1}{\mu-\Lambda\beta} - \frac{1}{\mu}\right), \\
& \bar{V}(1-\alpha) - c\left[\frac{1}{\mu-\Lambda\beta} + \frac{1}{\mu(1-\kappa\alpha)}\right] - P = 0.
\end{aligned}$$

B.1.4. Referral Strategy (iv): $(r^B, r^R) = (\kappa,1)$.

$$\begin{aligned}
\max_{P,\alpha,\beta,\kappa} \quad & P\Lambda\beta\left[1 + \frac{\kappa\alpha}{1-\alpha}\right] \\
\text{s.t.} \quad & \bar{V}(1-\beta) - c_r - c[\alpha W_1^B(\alpha,\beta,\kappa) + (1-\alpha)W_2^B(\alpha,\beta,\kappa)] - P = 0, \\
& c_r = c\alpha[W_2^B(\alpha,\beta,\kappa) - W_1^B(\alpha,\beta,\kappa)], \\
& \bar{V}(1-\alpha) - c_r - c[\alpha W_1^R(\alpha,\beta,\kappa) + (1-\alpha)W_2^R(\alpha,\beta,\kappa)] - P = 0,
\end{aligned}$$

where $W_i^\chi(\alpha,\beta) = \omega_i^\chi(\Lambda\beta,\kappa\alpha,\alpha), \quad i = 1,2, \ \chi \in \{B,R\}$.

B.2. Numerical Studies of Price Adjustments and Throughput Changes

Tables B.1 and B.2 report the numerical study of the firm's percentage price adjustment in the referral reward program and referral priority program, respectively. In both programs, the firm may either increase or decrease the price (the effective net price in the case of the referral reward program). Specifically, the firm tends to increase the price when the maximum service valuation $\bar{V}$ is high and the base market size Λ is intermediately large. Of course, when Λ gets too large, both programs would have no referrals and revert to FIFO, and there would not be any price adjustment (see Propositions 6 and 7).

Tables B.3 and B.4 report the numerical study of the firm's percentage throughput change in the referral reward program and referral priority program relative to FIFO, respectively. In both programs, the system

Table B.1 **Percentage change in price (%) of the referral reward program relative to the non-referral FIFO benchmark.**

Λ	$\bar{V}=5$	$\bar{V}=7.5$	$\bar{V}=10$	$\bar{V}=12.5$	$\bar{V}=15$	$\bar{V}=17.5$	$\bar{V}=20$
0.1	-19.64	-22.42	-24.96	-27.10	-28.96	-30.58	-32.03
0.3	-13.83	-17.38	-18.92	-20.30	-21.54	-22.63	-23.60
0.5	-4.85	-13.11	-13.67	-14.31	-14.94	-15.52	-16.08
0.7	0.00	-9.72	-9.46	-9.44	-9.53	-9.65	-9.80
0.9	0.00	-7.25	-6.32	-5.72	-5.36	-5.09	-4.90
1.1	0.00	-4.98	-4.17	-3.18	-2.44	-1.86	-1.38
1.3	0.00	-2.57	-2.82	-1.58	-0.61	0.21	0.85
1.5	0.00	-0.59	-2.04	-0.67	0.41	1.32	2.06
1.7	0.00	0.00	-1.34	-0.28	0.83	1.75	2.54
1.9	0.00	0.00	-0.74	-0.13	0.92	1.84	2.59
2.1	0.00	0.00	-0.20	-0.15	0.84	1.68	2.41
2.3	0.00	0.00	0.00	-0.13	0.65	1.47	2.12
2.5	0.02	0.00	0.00	-0.07	0.37	1.18	1.80
2.7	0.00	0.00	0.00	0.00	0.17	0.83	1.48
2.9	0.00	0.00	0.00	0.00	0.04	0.48	1.15
3.1	0.00	0.00	0.00	0.00	0.00	0.22	0.73

$c=1,\ \mu=1,\ c_r=0.2.$

Table B.2 **Percentage change in price (%) of the referral priority program relative to the non-referral FIFO benchmark.**

Λ	$\bar{V}=5$	$\bar{V}=7.5$	$\bar{V}=10$	$\bar{V}=12.5$	$\bar{V}=15$	$\bar{V}=17.5$	$\bar{V}=20$
0.1	0.00	0.00	0.00	-70.85	-62.00	-55.77	-51.16
0.3	0.00	0.00	-35.50	-24.50	-21.64	-22.74	-23.72
0.5	0.00	-37.17	-18.14	-14.46	-15.10	-15.69	-16.23
0.7	0.00	-28.63	-9.61	-9.60	-9.69	-9.82	-9.96
0.9	0.00	-24.51	-6.47	-5.91	-5.53	-5.25	-5.05
1.1	0.00	0.00	-4.32	-3.34	-2.59	-2.00	-1.52
1.3	0.00	0.00	-2.96	-1.72	-0.75	0.06	0.73
1.5	0.00	0.00	-2.57	-0.82	0.27	1.18	1.95
1.7	0.00	0.00	-3.75	-0.40	0.72	1.65	2.45
1.9	0.00	0.00	0.00	-0.26	0.83	1.73	2.51
2.1	0.00	0.00	0.00	-0.29	0.75	1.61	2.34
2.3	0.00	0.00	0.00	-0.39	0.58	1.38	2.06
2.5	0.00	0.00	0.00	4.96	0.38	1.12	1.75
2.7	0.00	0.00	0.00	3.79	0.17	0.86	1.44
2.9	0.00	0.00	0.00	2.54	-0.03	0.61	1.15
3.1	0.00	0.00	0.00	1.24	-0.21	0.39	0.88

$c=1,\ \mu=1,\ c_r=0.2.$

throughput may increase or decrease. Combining these two tables with Tables B.1 and B.2, we note that when the maximum service valuation $\bar{V}$ is high and the base market size is intermediately low, the firm jointly increases the price and throughput in both programs. One the other hand, the system throughput tends to decrease when the maximum service valuation $\bar{V}$ is high and the base market size is intermediately large. In the referral reward program in Table B.3, the firm's profit cannot be lower than that in the FIFO benchmark, so in all those cases when the throughput decreases, the firm earns a higher profit by raising

the price (more so than the decline in the system throughput). This is somewhat counter-intuitive because one would expect referrals to boost demand, but here, the firm leverages the referral program to charge a higher price and dampen demand. In the referral priority program in Table B.4, the firm's profit could be lower than that in the FIFO benchmark. However, in those italic cells, we find a similar phenomenon: the system throughput becomes lower in the referral priority program while the firm's profit is improved.

Table B.3 Percentage change in throughput (%) of the referral reward program relative to the non-referral FIFO benchmark.

Λ	$\bar{V}=5$	$\bar{V}=7.5$	$\bar{V}=10$	$\bar{V}=12.5$	$\bar{V}=15$	$\bar{V}=17.5$	$\bar{V}=20$
0.1	32.49	63.45	87.92	108.41	126.30	142.23	156.71
0.3	18.66	44.60	62.26	76.75	89.12	99.95	109.59
0.5	5.38	30.53	43.15	53.35	61.93	69.34	75.90
0.7	0.00	20.27	29.33	36.50	42.46	47.57	52.03
0.9	0.00	13.06	19.50	24.49	28.64	32.14	35.19
1.1	0.00	7.47	12.68	16.17	19.00	21.38	23.43
1.3	0.00	3.27	8.03	10.47	12.41	13.99	15.38
1.5	0.00	0.63	4.88	6.62	7.96	9.04	9.98
1.7	0.00	0.00	2.58	4.06	5.02	5.78	6.40
1.9	0.00	0.00	1.10	2.30	3.03	3.57	4.03
2.1	0.00	0.00	0.22	1.19	1.69	2.11	2.44
2.3	0.00	0.00	0.00	0.50	0.79	1.09	1.36
2.5	0.00	0.00	0.00	0.13	0.32	0.42	0.63
2.7	0.00	0.00	0.00	0.00	0.08	0.04	0.11
2.9	0.00	0.00	0.00	0.00	-0.01	-0.09	-0.21
3.1	0.00	0.00	0.00	0.00	0.00	-0.10	-0.25

$c=1,\ \mu=1,\ c_r=0.2.$

Table B.4 Percentage change in throughput (%) of the referral priority program relative to the non-referral FIFO benchmark.

Λ	$\bar{V}=5$	$\bar{V}=7.5$	$\bar{V}=10$	$\bar{V}=12.5$	$\bar{V}=15$	$\bar{V}=17.5$	$\bar{V}=20$
0.1	0.00	0.00	0.00	261.91	250.97	243.28	237.56
0.3	0.00	0.00	96.70	86.18	89.41	100.30	109.99
0.5	0.00	67.08	50.60	53.66	62.31	69.78	76.34
0.7	0.00	45.15	29.54	36.79	42.81	47.95	52.43
0.9	0.00	33.50	19.72	24.80	28.96	32.48	35.53
1.1	0.00	0.00	12.89	16.43	19.29	21.67	23.72
1.3	0.00	0.00	8.23	10.71	12.66	14.27	15.63
1.5	0.00	0.00	5.50	6.85	8.20	9.29	10.20
1.7	0.00	0.00	5.00	4.25	5.22	5.97	6.59
1.9	0.00	0.00	0.00	2.51	3.21	3.76	4.20
2.1	0.00	0.00	0.00	1.32	1.86	2.27	2.59
2.3	0.00	0.00	0.00	0.51	0.93	1.25	1.50
2.5	0.00	0.00	0.00	-5.17	0.29	0.55	0.75
2.7	0.00	0.00	0.00	-3.93	*-0.15*	0.06	0.23
2.9	0.00	0.00	0.00	-2.61	-0.46	*-0.28*	*-0.14*
3.1	0.00	0.00	0.00	-1.25	-0.68	-0.53	-0.40

$c=1,\ \mu=1,\ c_r=0.2.$