THE UNIVERSITY OF CHICAGO

INNOVATIVE MECHANISMS FOR MANAGING CUSTOMER ACCESS TO CONGESTED SERVICES

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE UNIVERSITY OF CHICAGO
BOOTH SCHOOL OF BUSINESS
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

BY
LUYI YANG

CHICAGO, ILLINOIS
JUNE 2017

Nobody goes there anymore. It's too crowded. — Yogi Berra

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost, I am deeply indebted to my advisor, Professor Laurens Debo, for his visionary advice in both academic and personal terms, unreserved encouragement on both life and research, and above all, immense support, both mentally and financially. Laurens left the University of Chicago for Dartmouth College starting from the third year of my PhD study; yet this departure has not at all detracted from his helpfulness as an advisor. This dissertation would not be completed without his tireless mentorship. His strong work ethics have influenced my character and will guide me in developing as a scholar in years to come.

I am also grateful to my co-advisor, Professor Varun Gupta. I benefited greatly from his expertise on queueing theory and his attentiveness to details in writing. He always engaged me in in-depth thinking about the dependency of qualitative insights from a model on its technical constructs. Thanks to his keen insight into the subtleties of the English language, I became a more articulate writer. His suggestions on diction improved the exposition of this dissertation and my writing skills in general.

My great gratitude also goes to other members of my dissertation committee: Professors John Birge, René Caldentey and Pradeep Chintagunta. Their comments and suggestions have significantly improved the quality of the dissertation. In particular, the positioning and research directions of Chapters 2 and 3 have benefited much from their wisdom.

I would like to express my special thanks to many practitioners with whom I have had long conversations about industry practice and external academics who provided constructive feedback for my dissertation. The research question in Chapter 1 was motivated by Chipeng Liang, a Chicago Booth alumnus, and his co-founded mobile technology startup Smartline. Professors Philipp Afèche and Leon Yang Chu also offered valuable suggestions to improve Chapter 1. Conversations with Professor Anita Tucker helped me better position Chapter 2 in the healthcare context. Chapter 3 benefited from discussions with Miles Wellesley from Robinhood Market and Justin McNally from Waitlisted.co. Professors Refael Hassin, Shiliang Cui, and Dongyuan Zhan have proffered their respective suggestions for multiple

chapters of this dissertation.

The University of Chicago provided me with a superb work environment for conducting interdisciplinary research in this dissertation. Courses I took in the Booth School of Business, the Department of Economics, and the Department of Statistics have laid a solid knowledge foundation for my dissertation research. I would like to particularly extend my appreciation to Professors Dan Adelman, Barış Ata, Rod Parker, and my advisor Laurens for stewarding the up-and-coming doctoral program of management science and operations management (MS/OM) as area advisors at various stages of my PhD career. I was humbled by the excellent curricula of the PhD program, namely Linear Programming with Kipp Martin, Dynamic Programming with Sunil Kumar, Revenue Management with Bora Keskin, Healthcare Operations with Don Eisenstein and Burhan Sandıkçı, Infinite-dimensional Optimization with Chris Ryan, Networks with Ozan Candogan, among others.

My time spent with friends and colleagues at the University of Chicago has been a cherished period of my life. I enjoyed waxing philosophical with classmates from the MS/OM PhD program. In particular, I would like to thank Adam Schultz for the many chats we had about all facets of life and for helping me pick up various idioms and expressions. I was fortunate to have Yifan Feng, Alex Langerman and Marko Rojnica as my collaborators, who shared with me their unique perspectives on research. I am grateful to Guanhao Feng, Yuan Mei, Yinan Su, and Hanzhe Zhang for organizing the Chicago Chinese Social Research Group, in which I presented much of my dissertation work. In addition, I owe a debt to Danna Zhang and Yuefeng Han for bugging them with many probability and statistics problems that arose in coursework and research.

Last but not least, I would like to thank my parents and my significant other for their greatest care, love and support throughout my PhD journey.

ABSTRACT

While goods can be produced and stocked, services are not delivered until customers encounter service providers. To the extent that providers are often capacitated, delays are commonly experienced by customers in their access to services. Advances in information and mobile technologies have spawned novel mechanisms for managing waiting customers. These mechanisms emerge in public and private sectors, in established service operations as well as startup companies. My dissertation develops theoretical models to understand (1) how individual customers, as rational economic agents, respond to these mechanisms; (2) what are the consequences and implications for the aggregate system performance; (3) how to design these mechanisms most effectively; (4) how these new mechanisms compare with existing ones.

Chapter 1 studies the optimal mechanism design of a two-sided marketplace where customers in a queue consensually *trade* their waiting spots. If a customer ever moves back in the queue, she will receive an appropriate monetary compensation. Customers can always decide not to participate in trading and retain their positions as if they are being served in the first-in, first-out (FIFO) queue discipline. The time-trading mechanism has the best of both worlds because it respects customers' property rights over their waiting positions as in FIFO and improves efficiency as in priority queues that serve customers with higher waiting costs. Chapter 1 designs the optimal mechanisms for the social planner, the service provider, and an intermediary who might mediate the trading platform. Both the social planner's and the service provider's optimal mechanisms involve a flat admission fee and an auction that implements strict priority. If a revenue-maximizing intermediary operates the trading platform, it should charge a trade participation fee and implement an auction with some trade restrictions. Therefore, customers are not strictly prioritized. However, relative to a FIFO system, the intermediary delivers value to the social planner by improving efficiency, and to the service provider by increasing its revenue.

Chapter 2 examines a setting in which customers looking for service providers face *search*

frictions and service providers vary in quality and availability. To understand customers' search behavior when they are confronted with a large collection of vertically differentiated, congested service providers, I build a model in which arriving customers conduct costly sequential search to resolve uncertainty about service providers' quality and queue length and select one according to an optimal stopping rule. Customers search due, in part, to variations in waiting time across service providers, which, in turn, are determined by the search behavior of customers. Thus, an equilibrium emerges. I characterize customers' equilibrium search/join behavior in a mean field model as the number of service providers grows large. I find that reducing either the search cost or customer arrival rate may not improve customer welfare and may strictly increase the average waiting time in the system as customers substitute toward high-quality service providers. Moreover, with lower search costs, the improved quality obtained by customers may not make up for the prolonged wait, therefore degrading the average search reward. Chapter 2 discusses policy implications of the results in the context of public surgical waits.

Chapter 3 investigates the *referral* priority program, an emerging business practice adopted by a growing number of technology companies that manage a waitlist of customers, which 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, the queueing-game-theoretic analysis in Chapter 3 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. I also compare the referral priority program with the referral reward program when the service provider optimally sets the admission price. I find that under a small base market size, the referral reward program would encourage referrals using monetary incentives. Numerical studies suggest the referral priority program is more profitable than the referral reward program when the base market size is intermediately small.

CHAPTER 1

TRADING TIME IN A CONGESTED ENVIRONMENT

1.1 Introduction

A queue is a place (physical or virtual) where individuals meet for a certain period of time to obtain sequential access to a scarce resource (such as a restaurant table, an event ticket, a newly released product). As such, a queue constitutes a "miniature social system" in which the underlying fabric that ties individuals to society also guides the relationships between those in a queue (Mann 1969). Rather than letting the question of who should access the resource first degenerate into a brawl, the behavioral protocol of queueing collectively endorses the notion of "property rights" (Gray 2009), a fundamental part of the social fabric. In particular, an individual's *position* in a queue is considered by its occupant as her *property* that she temporarily *owns*. Tampering with one's position in a queue amounts to taking away someone's property and may be met with strong objection: Any attempt to cut in line may be disapproved since this infringes on the "bumped" customers' perceived property rights over their waiting positions. This is one of the reasons why the first-in, first-out (FIFO) queue discipline is predominant in many services systems. The FIFO rule ensures "a direct correspondence between inputs (time spent waiting) and outcomes (preferential service)" and thus manifests a basic principle of distributive justice (Mann 1969). However, the FIFO rule does not accommodate heterogeneity in time-sensitivity of the queue occupants. The system would be more efficient if more time-sensitive customers jump ahead and get served faster. To that end, service providers often sell priorities to customers. For instance, EE, one of the largest telecommunications companies in the UK, launched a new service feature called "Priority Answer" that allowed customers to pay £0.50 to jump the queue for a service call. This new feature soon created a huge uproar and irked many customers who complained they were not being treated fairly (see e.g., BBC News Business 2014).

What goes awry with Priority Answer is that the proceeds go to the service provider,

yet a longer wait is inflicted on the non-paying customers. The misalignment would be resolved if the monetary transfer were among customers themselves: impatient customers may be willing to pay to acquire the position of less impatient customers who are potentially willing to give away their spots for monetary rewards. This calls for a two-sided marketplace where customers consensually *trade* their waiting spots. Such a marketplace enables waiting customers to voluntarily swap positions at mutually agreed prices. Since such swaps do not influence the positions of any other customers on the wait list, no customers are forcibly pushed back without being compensated. Thus, customers can have the best of both worlds: their proprietary entitlements to waiting positions are adequately acknowledged as in the FIFO system; and their diverse priority requests are aptly reconciled, improving system efficiency.

One impediment to such a two-sided market is its practical organization: as service progresses and new customers arrive, the dynamic nature of the queue makes it difficult to negotiate prices and reorganize the queue continuously in real time. Fortunately, advances in mobile technologies make such marketplaces more implementable than ever before. In contrast to clamoring crowds bargaining for a better deal, customers can trade on a mobile platform without wreaking too much havoc in a queue. This is even more appealing when customers use the mobile app to join queues remotely so there will be little scope for chaotic transactions.

In this chapter, we study how trading in a queue can be organized by simple auctions in an environment where customers are privately informed about their waiting costs. We design the optimal mechanisms from three different perspectives: social welfare, the service provider's revenue, and the revenue of the trading platform (which we refer to as the intermediary) that mediates trading. While the first two perspectives are common in the queueing literature, they implicitly rely on the assumption that the trading platform is and can be managed by the service provider, which may not necessarily be true in practice. Instead, the service provider may be inclined to delegate the trading platform to an intermediary

for technological reasons and reputational concerns. First and foremost, the infrastructure that facilitates trade hinges on technology (e.g., mobile apps) that typically falls beyond the expertise of the service provider. Therefore, if a specialized intermediary is responsible for developing, deploying and maintaining the platform on behalf of the service provider, the service provider will not be distracted from its core competencies. In the restaurant industry, for example, dining reservation platforms (intermediaries) are typically not fully integrated with restaurants (service providers): examples include OpenTable which charges restaurants for each reservation (Stross 2010), and a similar dining app, Reserve, which alternatively charges customers for each booking (DeAmicis 2014). Second, if the service provider were to operate and conceivably profiteer from a resale market of waiting positions (either directly by collecting fees for trading or indirectly via surcharges in service fees), there might be a backlash from customers given the sensitive nature of queue-jumping (as in the case of Priority Answer). To the extent that this results in a loss of goodwill, the service provider would rather be detached from the trading platform and leave it to a third-party intermediary to arbitrate swaps of waiting positions. This begs the question of what is the optimal mechanism to collect fees from trading customers for an intermediary, who has the potential of raising sizable revenues once the technology is scaled up.

We first study a simple baseline auction scheme as a building block where each arriving customer first decides whether to join the queue or balk, and if she joins, she further decides whether to participate in trading. If she chooses to trade, she submits a sealed bid. This bid represents her selling price and maximum buying price per unit of time traded. She buys waiting positions from any customer who bids strictly lower and sells positions to any customer who bids strictly higher. In each transaction, the monetary transfer is equal to the seller's price multiplied by the expected waiting time exchanged. Any two customers with a tie in bids do not trade and are served according to the FIFO discipline. If a customer chooses not to trade, she retains her waiting position as if in a FIFO system and her expected waiting time is not affected by trading among other customers.

We find that in this baseline auction, all joining customers voluntarily participate in trading in equilibrium and customers prioritize themselves in decreasing order of their waiting cost, achieving the socially optimal service order, a special case of the $c\mu$ rule (Cox and Smith 1961) when customers have homogeneous service requirements. However, this mechanism does not maximize social welfare because customers end up overjoining. To recover the socially optimal arrival rate, an admission fee should be administered that applies to all joining customers. If the goal is to maximize the service provider's revenue, the same auction can be run but with a higher admission fee.

By contrast, an intermediary operating the trading platform charges a trade participation fee to generate revenue. If it subsequently runs the baseline auction, there will be an upper bound on the trade participation fee beyond which some moderately time-sensitive joining customers choose not to trade. To overcome this limit on the volume of trading customers, we propose an augmented auction with, somewhat counter-intuitively, two prices that put a restriction on trades. The only difference from the original auction is that if both customers submit bids in between the two prices, they are prohibited from trading with each other. We show that this auction format is optimal for the intermediary, and that the optimal trade participation fee is, in fact, higher than the aforementioned upper bound while all joining customers still voluntarily participate in trading. The auction results in a partial pooling equilibrium where some customers with different waiting costs bid the same amount and expect the same waiting time as if they are served according to the FIFO rule, but since they still trade, other customers expect the waiting time as if they are served efficiently according to the $c\mu$ rule. In addition, allowing the intervention of a revenue-maximizing intermediary, the service provider always earns a higher revenue than it would in a FIFO system. Therefore, the intermediary's value proposition is twofold: (i) improving system efficiency, and (ii) increasing the service provider's revenue.

The contributions of our work to the literature are twofold. We contribute to the queueing literature by investigating how queue positions can be reorganized via voluntary trading

among customers in various auctions that maximize social welfare, the service provider's revenue, and a trade-mediating intermediary's revenue, respectively. We contribute to the literature on mechanism design with an operations management application where the allocation of a special type of good, expected waiting time, depends on the operational characteristics of a queue.

The remainder of the chapter is organized as follows. In §1.2, we review the related literature. We set up the model in §1.3. In §1.4, we propose the baseline auction, which is a building block for the optimal mechanisms to maximize social welfare and the service provider's revenue, respectively, in §1.5. In §1.6, we propose the augmented auction for maximizing the intermediary's revenue and then study its impact on the service provider's pricing decision and revenue. We conclude the chapter and discuss future research directions in §1.7.

1.2 Related Literature

Our study bridges the economics literature on mechanism design in trading, and the operations literature on priority queueing systems. It is also related to the sociology/psychology/law literature on priority rights and social norms in queues.

1.2.1 Mechanism Design in Trading

Bilateral Trading. Mechanism design techniques (e.g., Myerson 1981) are applied to many important trading problems with private information. One seminal result is the impossibility theorem in bilateral trading proved by Myerson and Satterthwaite (1983). They find that no efficient mechanism can be Bayesian-Nash incentive compatible, individually rational, and does not run a deficit. This runs counter to the celebrated Coase theorem (Coase 1960) which states that a free market leads to an efficient outcome regardless of the initial property allocation. The drivers of Myerson and Satterthwaite (1983) are the presence of

property rights and private information about valuation. Chatterjee and Samuelson (1983) and Satterthwaite and Williams (1989) study the same setting in a double auction scheme.

Unlike Myerson and Satterthwaite (1983), we show in our queueing context that trading can achieve efficient ordering, but similar to their result, it does not maximize social welfare[1] due to a higher arrival rate than is socially optimal. Myerson and Satterthwaite (1983) also study a setting where trade is mediated by an intermediary and show that it is optimal for the intermediary to restrict some trading as a monopolist. However, they do not discuss the implementation of the optimal mechanism, e.g., a fee-setting double auction. We also consider the intermediary's problem with customers' endogenous queue-joining decisions, and we extend the optimality of trade restriction to our queueing environment. Moreover, we show that a simple, but, subtle, three-parameter auction can implement the revenue maximizing trading mechanism for the intermediary. We demonstrate that the intermediary benefits the service provider who would otherwise manage a FIFO queue. In the bilateral trading models described above, the roles of the seller and the buyer are fixed, whereas in our queueing environment, today's buyers can become tomorrow's sellers who resell their waiting spots to newly arrived customers.

Partnership Dissolution. Cramton et al. (1987) consider a partnership where each player is endowed with a share of a divisible good to be traded and is privately informed about their valuation. In contrast to Myerson and Satterthwaite (1983), they show that partnership dissolution can be ex-post efficient (i.e., the player who values the good most gets the entire share) provided the initial ownership is not too imbalanced (e.g., equal shares). We obtain a similar result of full efficiency in our queueing environment when customers initially expect a FIFO waiting time. Kittsteiner (2003) considers interdependent valuation in partnership dissolution. They extend the strategy space to incorporate vetoing against dissolution. Al-

1. In the setting of Myerson and Satterthwaite (1983), ex-post efficiency is synonymous with maximizing social welfare as the number of traders is fixed (one buyer and one seller), but since we also consider customers' endogenous joining decisions, maximizing social welfare requires both the socially optimal service order (efficiency) and arrival rate.

lowing customers to veto against trading is also a key feature in our model to reflect the ownership of one's queue position.

1.2.2 Priority Auction and Pricing

Priority Auction. Kleinrock (1967) starts the stream on priority auctions/queue bribery. In his work, each arriving customer bids for priority. However, Kleinrock's payment functions are exogenously given. Lui (1985) and Glazer and Hassin (1986) extend Kleinrock's model by deriving customers' endogenous bid functions in equilibrium. They show that the $c\mu$ rule can be achieved in this priority auction. Hassin (1995) shows that the bidding mechanism in the priority auction is self-regulating in that it achieves both the socially optimal service order and arrival rate. Afèche and Mendelson (2004) consider a generalized delay cost structure in the priority auction. In Kittsteiner and Moldovanu (2005), customers possess private information on job processing time.

In these papers, all the monetary payments are made to the service provider and there is no monetary transfer among customers. Customers' property rights over their waiting positions are not taken into account, and therefore, if a new customer bids higher than an existing customer, the existing customer must give way to the new customers without receiving any compensation. Moreover, if a customer joins the system, she must participate in the auction. Since our model acknowledges customers' ownership of queue positions (i.e. the property rights), customers are not forced to participate in trading even after they join the system.

Priority Pricing. This literature is pioneered by Mendelson and Whang (1990), who study the incentive-compatible socially optimal priority pricing scheme. From the mechanism design perspective, priority pricing is de facto tantamount to a direct revelation mechanism where customers truthfully report types. Recently, Afèche (2013) and Afèche and Pavlin (2016) study this problem from a revenue-maximizing service provider's perspective. Both papers use the achievable region method due to Coffman and Mitrani (1980) as their solution

approach rather than pre-specifying a particular scheduling policy. Our work also applies this solution methodology, but we find simple auctions to implement the optimal mechanisms. In addition, both Afèche and Pavlin (2016) and Katta and Sethuraman (2005) find it optimal sometimes to pool some customers into a single FIFO class. The pooling result in Afèche and Pavlin (2016) hinges on the customer type ranking being lead-time dependent, whereas in Katta and Sethuraman (2005) it is a consequence of the non-monotonicity of the hazard rate of customer type distribution. Our model does not have such features, yet the intermediary's optimal mechanism also requires pooling. It is driven, instead, by the intermediary's incentive to restrict trading as a monopolist when customers claim property rights in the queue. Besides, the pooling feature in our mechanism is robust to all model primitives under consideration, while in the aforementioned papers, pooling is a special case that occurs under a specific set of parameters.

We refer the reader to Hassin and Haviv (2003) for an extensive survey of queueing models with strategic customers.

1.2.3 Property Rights and Social Norms in Queues

While the importance of social norms in queues has long been recognized in the sociology and psychology literature (e.g., Mann 1969 and Schwartz 1975), Gray (2009) is among the first attempts to formalize the concept of property in queues and explore the proprietary significance of queueing protocols. He contends that the (mal)practice of queue bribery in Kleinrock (1967) is a subversion of queueing norms as it is a form of queue-jumping that generates feelings of social injustice. By comparison, he reckons that trading waiting spots by way of direct substitution imposes no externalities for other members in the queue, and such transfers comply with the normative code of the queue. Oberholzer-Gee (2006) asserts that there is a natural incentive for Pareto-improving trades since the arriving order of individuals does not reflect their opportunity cost of time. Allon and Hanany (2012) build a game-theoretic model to show that cutting in line can be sustained under rational decision

making in repeated interactions, through which a social norm can develop. While they focus on situations without monetary payments, our work complements their research by allowing, via technology, monetary exchanges to sell and purchase waiting spots. We show that with technology and money, efficiency can be increased. This is important because technology can be easily managed and scaled, whereas social norms might be more difficult to replicate or modify.

1.2.4 Trading in a Queue

Like this chapter, several papers address the problem of trading waiting positions in a queue. Rosenblum (1992) assumes that customers' waiting costs are public information in their trading model and that future values of transactions are ignored. In our model, customers privately observe their own waiting costs and take into account the expected values of future transactions when they determine their bids in the auction. Gershkov and Schweinzer (2010) formulate a mechanism design problem of rescheduling a fixed number of players in a clearing system where there is no arrival process and trading is completed before service starts. Since all customers are present at time zero, it is not clear how the initial property rights are formed, so they study different initial allocations and show that an efficient mechanism can be implemented if the initial schedule is random ordering but not if it is deterministic like FIFO. El Hajia and Onderstal (2015) experimentally examine how human subjects trade in a queueing environment similar to Gershkov and Schweinzer (2010). They provide evidence that organizing such a time-trading market can achieve a nontrivial amount of efficiency gains. Our model incorporates the operational dynamics of a queueing system where the order of arrivals naturally gives rise to the initial allocation. It allows us to study how trading impacts customers' endogenous queue-joining behaviors. While Gershkov and Schweinzer (2010) and El Hajia and Onderstal (2015) mostly focus on the efficiency of the time-trading mechanisms, our work also incorporates the perspectives of the revenue maximizing service provider and intermediary, and finds the optimal mechanisms for each of the performance

measures.

1.3 Model Setup

Consider a congested service facility, modeled as an $M/M/1$ queueing system, that faces a population of delay-sensitive customers.

Customer Base. Customers arrive at the system according to an exogenous Poisson process with rate Λ (market size). Each customer requests one unit of service. The service times are i.i.d. samples from an exponential distribution with mean $1/\mu$. Let $\rho \doteq \Lambda/\mu$.

Customer Valuation. Customers have a common valuation V for service. For a customer with delay cost rate c, who experiences waiting time w, defined as the entire duration in the system, and money m after receipt of service, her utility is $V - c \cdot w + m$. For simplicity, we normalize initial money wealth for all customers at zero, but assume that they are not budget-constrained, so $m > 0$ means a customer is a net receiver; $m < 0$, a net payer. Customers differ in their delay cost rate c. Each customer's delay cost per unit time is an i.i.d. draw from a continuous distribution with a strictly increasing cumulative distribution function F and a finite, strictly positive and continuously differentiable probability density function f over the support $c \in \Xi \triangleq [\underline{c}, \overline{c}]$ and $0 \leq \underline{c} < \overline{c} < \infty$. Customers are risk-neutral and expected utility maximizers. To exclude the case where no customers have a positive net value even if served immediately, we make Assumption 1.1.

Assumption 1.1. $V > \underline{c}/\mu$.

Upon arrival, customers decide whether to join the service facility to obtain service, or balk. In case they do not join, they obtain the reservation utility, which we normalize to zero. If they join, customers do not renege subsequently.

Information Structure. The inter-arrival time distribution, the service time distribution, the delay cost distribution f and the service value V are common knowledge. The type of each individual customer (delay cost rates c) is her private information. As commonly

assumed in the literature (e.g., Kleinrock 1967), customers do not observe the system state upon arrival but can estimate the expected waiting time and the expected monetary transfer.

1.4 Baseline Auction

In this section, we study an auction-based trading mechanism that is budget-balanced among customers: all monetary transfers are internal within customers. This auction is the building block for subsequent results about the social planner, service provider and intermediary in §1.5 and §1.6.

1.4.1 Trading Rules

Auction format: In the baseline auction, upon arrival, a customer decides whether to join the queue or not. If the customer does not join, she earns a reservation utility of 0. If the customer joins, she submits a sealed bid b that can either be "No," or a price for one unit of time. We allow customers to bid "No" to reflect that trading is voluntary and that customers can always preserve their FIFO property right. The bid b represents the least amount she wants to receive for expecting to wait one additional unit of time and also the greatest amount she is willing to pay for one unit of the expected waiting time reduced. The queue is reorganized in such a way that the arriving customer swaps positions consecutively with those who place bids strictly lower than hers. In each transaction, the customer who jumps ahead (the buyer) compensates the one who moves back (the seller) by the seller's bid price times the expected waiting time exchanged. The existing customers who submitted bids strictly higher or those who submitted "No" are not affected in their waiting position. Nor are customers who bid the same amount as the buyer. Any customers with equal bids are served FIFO amongst themselves. Note that this auction follows a "pay-as-you-overtake" paradigm, since customers' realized payment as buyers depends on the actual number of customers they overtake. For simplicity, trading is instantaneous (transactions do not take

any time) and preemptive-resume (customers at the server can suspend their service and sell their spot; service is resumed when this customer reaches the server again). Customers submit a bid before observing the queue length and commit to the bid throughout their stay in the system[2].

Illustration 1: Consider an arriving customer who joins and participates in trading by submitting a price b'. Assume that there are four other existing customers in the system. Among them, the first, the second and the fourth customer participate in trading with bids b_1, b_2 and b_4 (with $b_1 \geq b' > b_2 > b_4$), respectively. The third bids "No" and thus does not participate in trading. Thus, before the new arrival, the system can be represented by (b_1, b_2, F, b_4), where F stands for a FIFO customer who bids "No". Adding the arriving customer (customer 5) who bids b' to the tail of the queue, we have (b_1, b_2, F, b_4, b'), which is not (yet) ordered. Then, the auction swaps customer 5 and customer 4, yielding (b_1, b_2, F, b', b_4). Customer 5 makes a payment of b_4/μ to customer 4. Next, the auction swaps customer 5 and customer 2, yielding (b_1, b', F, b_2, b_4). Notice that the expected wait time of the FIFO customer does not change. Customer 5 makes a payment of $2b_2/\mu$ to customer 2 (because the latter moves back by two positions). Customer 5 does not swap positions with customer 1 since customer 1 bids weakly higher and they are served FIFO. Thus, the trading process is completed. The total payment customer 5 makes to the other customers is thus $(b_4 + 2b_2)/\mu$. Similarly, customer 5 expects a compensation of b' per unit of time if she ever moves back and swaps positions with other, later arriving customers who make a higher bid than b'.

1.4.2 Auction Equilibrium

Strategy: We focus on symmetric pure strategies specified by two functions; the joining function $J : \Xi \mapsto \{\text{join, balk}\}$ specifies which customer types join or balk, and the bid function $b : \{c | J(c) = \text{join}\} \mapsto \mathbb{R}_+ \cup \{\text{No}\}$ specifies the bid of each customer type (either

2. After submitting their bid, customers could see the queue length, but this would be technically irrelevant to the bidding game since the trading process goes on autopilot once bids are collected from customers.

a price for one unit of time or "No"). Thus the effective arrival rate to the system is $\lambda \triangleq \Lambda \int_{\underline{c}}^{\overline{c}} \mathbf{1}\{J(c) = \text{join}\} dF(c)$, where $\mathbf{1}\{X\}$ is the indicator function of condition X.

Waiting time and utility: Given the bid function $b(\cdot)$ and the joining function $J(\cdot)$, let $W : \mathbb{R}_+ \mapsto \mathbb{R}_+$ denote the mapping from a customer's bid to her expected waiting time. Since trading does not affect any joining customer who bids "No," it is immediate that these customers' expected waiting time is equal to the mean waiting time of an $M/M/1$ system: $W(\text{No}|b, J) = \frac{1}{\mu - \lambda}$. Note that this waiting time depends on the endogenously determined λ, the aggregate arrival rate of the system, and is not impacted by any individual, infinitesimal customer's action. Since customers submit their bid up front and make a commitment during their wait, they take into account all future transactions in the expected utility (note that this is one of the key distinctions from Rosenblum 1992). We assume that customers do not discount future payments. Let $P_p : \mathbb{R}_+ \mapsto \mathbb{R}_+$ be the function that maps a customer's bid to the total expected amount of money she *p*ays as a buyer upon arrival; and $P_r : \mathbb{R}_+ \mapsto \mathbb{R}_+$ maps a customer's bid to the total expected amount of money she *r*eceives as a seller during her stay in the system.

Thus, given $b(\cdot)$ and $J(\cdot)$, the expected utility of a joining customer of type c who bids β is

$$U(c, \beta|b, J) = \begin{cases} V - cW(\beta|b, J) - P_p(\beta|b, J) + P_r(\beta|b, J), & \beta \in \mathbb{R}_+ \\ V - \frac{c}{\mu - \lambda}, & \beta = \text{No}. \end{cases} \tag{1.1}$$

We fully specify in Appendix A.2 the expressions of $W(\beta|b, J)$, $P_p(\beta|b, J)$, $P_r(\beta|b, J)$ in equilibrium.

Customer Equilibrium: A symmetric pure-strategy Nash equilibrium is defined by the

following conditions:

$$b(c) \in \arg\max_{\beta \in \mathbb{R}_+ \cup \{\text{No}\}} U(c, \beta | b, J), \quad \forall c \in \{c | c \in \Xi,\ J(c) = \text{join}\} \tag{1.2a}$$

$$U(c, b(c) | b, J) \geq 0, \quad \forall c \in \{c | c \in \Xi,\ J(c) = \text{join}\} \tag{1.2b}$$

$$U(c, \beta | b, J) \leq 0, \quad \forall c \in \{c | c \in \Xi,\ J(c) = \text{balk}\}, \forall \beta \in \mathbb{R}_+ \cup \{\text{No}\}. \tag{1.2c}$$

Condition (1.2a) states that for all the joining customers, the best response of the equilibrium bid function should be itself. Condition (1.2b) ensures that all joining customers get nonnegative expected utility and (1.2c) specifies that the balking customers in equilibrium have no incentive to join the system since their expected utility would not turn positive regardless of what she bids.

An equilibrium is said to achieve *efficiency* or be an *efficient schedule* if $b(c)$ is strictly increasing in c whenever $J(c)$ =join. If this holds, customers are effectively prioritized by the $c\mu$ rule.

It is immediate that there is a trivial equilibrium: all joining customers submit "No". Thus, nobody participates in trading and customers are served FIFO. This equilibrium holds in all auction settings in this chapter. We analyze other equilibria that realize gains from trade. We indicate the equilibrium in the baseline auction by means of a superscript B.

Theorem 1.1 (Full Trading, Separating Equilibrium)**.** *Under the baseline auction, there exists an equilibrium in which*

(i) $J^B(c)$ =*join for* $c \in [\underline{c}, \tilde{c}]$ *(and balk otherwise) with* $\tilde{c} \leq \overline{c}$*, i.e.,* $\lambda^B = \Lambda F(\tilde{c})$*;*

(ii) the equilibrium bid function is strictly increasing and given by

$$b^B(c; \tilde{c}) = c + \frac{\int_c^{\tilde{c}} (F(\tilde{c}) - F(s))^2 W^e(s; \tilde{c}) ds}{(F(\tilde{c}) - F(c))^2 W^e(c; \tilde{c})}, \quad c \in [\underline{c}, \tilde{c}]$$

where $W^e(c; \tilde{c}) = \frac{1}{\mu[1-\rho(F(\tilde{c})-F(c))]^2}$ *is the time customer* c *expects to wait given* $\tilde{c}$*;*

(iii) the equilibrium expected utility of the joining customers, $U(c, b^B(c; \tilde{c}))$*, is convex de-*

creasing in c. Either $\tilde{c}$ *uniquely solves* $U(\tilde{c}, b^B(\tilde{c};\tilde{c})) = 0$ *or* $\tilde{c} = \overline{c}$ *if there is no solution.*

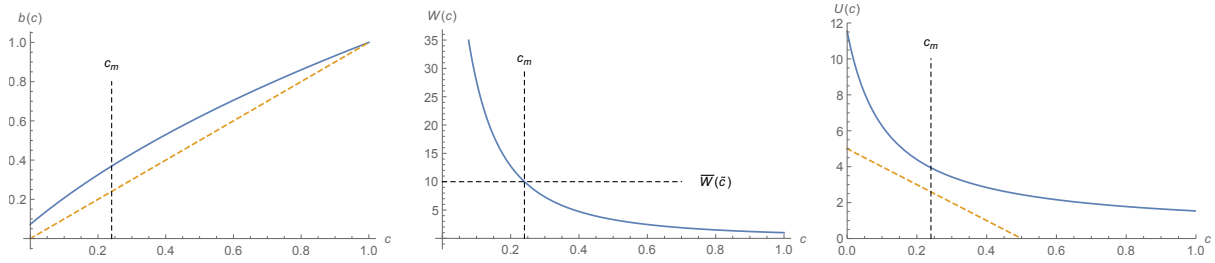
We illustrate Theorem 1.1 in Figure 1.1. Unless otherwise stated, we use the parameters in Table 1.1 for numerical illustrations throughout this chapter.

Table 1.1: Model primitives for numerical illustrations.

V	Λ	μ	F
5	0.9	1	$U[0,1]$

Theorem 1.1 suggests that customers follow a threshold policy in their joining decisions, and they balk if their waiting cost is high, i.e., c is greater than the cutoff value $\tilde{c}$. We henceforth use $\overline{W}(\tilde{c}) \triangleq \frac{1}{\mu - \Lambda F(\tilde{c})}$ to denote the expected FIFO waiting time. In this equilibrium, however, all the joining customers participate in trading. Most importantly, the equilibrium bid function is strictly increasing and the expected waiting time is strictly decreasing in c, implying that the budget-balanced baseline auction implements an efficient schedule: any two customers with different waiting costs trade waiting spots with one another so that customers are prioritized in decreasing order of their waiting cost. The resulting expected wait time is illustrated in Figure 1-(b).

Figure 1.1: The baseline auction.



(a) Biding function. (b) Waiting time. (c) Utility.

Note. The solid curves are the bidding function in (a), the expected waiting time in (b), and the expected utility under the auction. The dashed lines are a 45-degree line in (a), and the expected utility if customers are served FIFO under arrival rate λ^B in (c). In this example, $\tilde{c} = \overline{c} = 1$.

Achieving allocative efficiency via a budget-balanced trading mechanism under private information is a nontrivial result under individual rationality in trading (cf. Myerson and Satterthwaite 1983). As illustrated in Figure 1.1-(c), under this equilibrium, trading makes

all the joining customers better off, i.e., their expected utility exceeds the utility they would get if they bid "No" and wait FIFO. The intuition is the following. Prior to trading, all customers expect FIFO waiting time, so the initial waiting time allocation is the same. This is analogous to having equal shares before partnership dissolution (cf. Cramton et al. 1987). When trading starts, customers may buy from existing customers, and may also sell to future arriving customers. The countervailing incentives as both buyers and sellers offset each other, making an efficient schedule possible.

We also highlight that one favorable feature of the budget-balanced auction is that it has very simple rules that do not require knowledge of customer valuations like V and F. An efficient schedule is automatically achieved by customers themselves. While the auction, in principle, allows for an arriving customer to overtake a line-holder who chooses not to participate in trading (without influencing her expected waiting time, e.g., customer 5 overtakes the FIFO customer in Illustration 1), this case never happens in equilibrium as all customers trade, i.e., any customer who is overtaken gets compensated at an agreed-upon price in equilibrium.

As illustrated in Figure 1-(a), all the joining customers overbid, i.e., $b^B(c) > c$, except that customer $\tilde{c}$ bids truthfully, i.e., $\lim_{c\to\tilde{c}} b^B(c) = \tilde{c}$ ($\tilde{c} = \bar{c}$ in this example). This is a consequence of the auction rule that the seller's bid dictates the trading price in each transaction. Thus, customers have an incentive to inflate their bid as a seller to gain more revenue. An immediate implication of this bidding behavior is that in each trade, the buyer may be temporarily worse off. This happens if the buyer's waiting cost c_b is only slightly higher than the seller's c_s ($c_b > c_s$) and the seller's bid exceeds the buyer's waiting cost, i.e., $b^B(c_b) > b^B(c_s) > c_b > c_s$. In this case, the buyer pays more than she gains in waiting cost reduction, i.e., $b^B(c_s)\Delta W > c_b\Delta W$, where ΔW is the expected waiting time exchanged. Even though they incur losses on some trades, customers are still willing to participate in trading since they always gain as sellers, and also gain from some trades as buyers[3] so that

3. When overtaking customers who bid less than c_b.

the overall benefit of trading is still positive ex-ante.

As illustrated in Figure 1.1-(c), customers' expected utility is downward sloping in their waiting costs, implying that the more impatient a customer is, the less she gains from joining the system; whereas the convexity of the curve implies that customers with extreme waiting costs (either very high or very low) have the most gains from trading relative to being served FIFO whereas customers with medium waiting cost reap the least relative gains. Customers with very low waiting cost favor trading since they are most willing to sell spots for money, i.e., $P_p(b^B(\underline{c})) = 0$ and $P_r(b^B(\underline{c})) > 0$; while those with high waiting cost benefit from trading since they are most willing to pay to skip the line, i.e., $P_r(b^B(\tilde{c})) = 0$ and $P_p(b^B(\tilde{c})) > 0$. Both incentives are weak for customers in the middle. In particular, there is one type of customers, c_m, who expects exactly the same waiting time as if she were served FIFO, i.e., $W^e(c_m; \tilde{c}) = \overline{W}(\tilde{c})$. Because of the convexity of the utility curve, c_m is also the type of customers whose gain in trading relative to FIFO is the smallest. Still, she strictly prefers trading to FIFO since $P_r(b^B(c_m)) > P_p(b^B(c_m))$, i.e., the amount she expects to receive exceeds the amount she expects to pay. In a nutshell, two types of customers warrant special attention: the one with the least patience who would be the most sensitive to joining; and the one with moderate patience who would be the most sensitive to trading.

1.4.3 Discussion on the Auction Format

Before we proceed, three characteristics of the auction rule are worthy of discussion.

Single bid. Upon arrival, each customer is only requested to submit a single bid while an alternative auction would collect both a bid price (for purchase) and an ask price (for sales). Technically, solving for the equilibrium of the alternative auction requires jointly solving two linked differential equations of the bid function, $b(c)$ and ask function, $a(c)$. Moreover, the efficiency result would be lost since the two prices submitted by a given customer typically differ in equilibrium: a patient customer c_1 might fail to sell her position to an impatient customer c_2 ($c_2 > c_1$) if the former asks a higher price than the latter bids, i.e., $a(c_1) > b(c_2)$.

Trading price. In the present auction, the trading price is determined solely by the seller's price; the buyer's price only determines with whom one trades, but does not influence the price at which trade is executed. Alternatively, a common split-the-difference rule could be proposed where the trading price is a weighted average of the buyer's price and the seller's price (cf. Chatterjee and Samuelson 1983). This again creates a multiplicity of technical challenges: the underlying differential equation becomes more involved; there might be multiple equilibria; the conjectured bidding function might not be monotone. From a high-level standpoint, since now the buyer's price can influence the trading price, customers might have an incentive to shade their bid; theoretically, this might lead to negative bids in certain cases, which is more difficult to implement in practice.

Bid per unit time. As opposed to the typical priority auction (Kleinrock 1967) where customers pay their bid in a lump sum, our auction is such that the ex-post payment made by an arriving customer rests on the number of customers she overtakes. Since no queue length information is available when the bid is submitted, the strength of this "pay-as-you-overtake" approach is that how much one pays is precisely based on how much wait time one expects to reduce. By contrast, the typical priority auction suffers from a potential mismatch between time and money: a customer may pay a large (fixed) amount, only to find a short line skipped. On the other hand, the exact payment is unknown ex-ante in our auction, which may be unfavorable to real-world customers who are risk-averse or budget-constrained.

1.5 Social Welfare and Service Provider's Revenue

In this section, we study how social welfare and the service provider (SP)'s revenue can be maximized using the trading mechanism proposed in §1.4.

1.5.1 Social Optimization

Definition 1.1. *The maximum social welfare SW is determined by:*

$$SW = \max_{\tilde{c}\in\Xi:\Lambda F(\tilde{c})<\mu} \Lambda F(\tilde{c})V - \Lambda \int_{\underline{c}}^{\tilde{c}} cW^e(c;\tilde{c})dF(c). \tag{1.3}$$

The socially optimal arrival rate $\lambda^{SW} = \Lambda F(\tilde{c}^{SW})$, where $\tilde{c}^{SW}$ is the maximizer of (1.3).

Definition 1.1 formalizes the concept of the social optimum in a centralized system where the social planner can dictate the arrival rate and scheduling policy. First, it is socially optimal to serve customers with the smallest waiting costs for any arrival rate and scheduling policy; thus, the social optimum requires a threshold joining policy, which coincides with the equilibrium structure of the baseline auction. Second, for any arrival rate, it is socially optimal to prioritize customers by the $c\mu$ rule, which is achieved by the equilibrium structure of our trading mechanism. It is natural to ask whether the baseline auction as a decentralized mechanism implements the social optimum. Proposition 1.1 indicates the answer is negative in general.

Proposition 1.1. *$\lambda^B \geq \lambda^{SW}$ with equality if and only if $\lambda^B = \lambda^{SW} = \Lambda$. The social planner can achieve SW by running the baseline auction with an admission fee $p^{SW} = \int_{\underline{c}}^{\tilde{c}^{SW}} c\left[1 - \frac{F(c)}{F(\tilde{c}^{SW})}\right]\left[-\frac{\partial W^e(c;\tilde{c}^{SW})}{\partial c}\right] dc$.*

Although the baseline auction achieves the "right" service order (efficiency), it does not attain the socially optimal arrival rate in general: in particular, customers with high waiting cost who should otherwise balk in social optimum join the system under the trading mechanism. This runs counter to the well-known result for the typical priority auction as in Kleinrock (1967) which is shown to be self-regulating in both the arrival rate and service order (Hassin 1995). The problem with the trading mechanism is that unlike in the priority auction, customers do not fully internalize the negative externalities inflicted on others. They do pay for the cost imposed on the existing customers if they jump over them; in fact, they

are over-penalized in our auction since the trading price overstates the seller's waiting cost. However, they are not held accountable for the cost imposed on future arrivals; worse still, they can even earn rents on their waiting spots for future customers to buy. The inability to achieve the maximum social welfare is similarly found in the bilateral trading model in Myerson and Satterthwaite (1983), but takes a different form. There, maximizing social welfare is synonymous with achieving ex-post efficiency due to a fixed number of traders (one buyer and one seller). Their system is afflicted by the lack of ex-post efficiency, hence a loss in social welfare. Our queueing system attains efficiency for a given arrival rate, but customers' joining decisions are endogenous, precisely because of which, the system suffers from over-joining, again engendering a loss in social welfare. Fundamentally, this loss in social welfare is symptomatic of the presence of property rights: customers take the FIFO waiting time as their initial property and thus do not internalize the externalities inflicted on those who arrive later.

Like Naor (1969), an intuitive remedy to over-joining is to charge an admission fee p^{SW}. This fee can be interpreted as what the service provider charges for accessing the service facility, and thus it applies to all joining customers regardless of their trading decision. It is important to recognize that the admission fee only alters customers' joining incentives, but not their trading incentives since it decreases their utility if they trade just as much as it does the utility obtained from waiting FIFO. To the extent that all money flows are viewed as internal transfers, Proposition 1.1 shows that an appropriate admission fee can restore the social optimum. Charging a single admission fee and running the baseline auction for trading, this mechanism is *outcome equivalent* to the aforementioned priority auction that regulates both the arrival rate and service order (Hassin 1995), but customers' perception can be quite different. In our mechanism, the service provider charges a flat fee for admission and customers sort out the right service order by themselves through trading. Moreover, joining customers can also opt out of the auction and maintain their FIFO position, but it just so happens that they all voluntarily trade in equilibrium. The second and third

Table 1.2: Comparisons of different mechanisms.

	Baseline	Socially optimal	SP revenue maximizing	FIFO pricing
Social welfare	2.942	3.106	2.921	2.121
SP's revenue	0	2.121	2.381	2.011
Admission fee	0	2.592	3.439	3.505
Arrival rate	0.9	0.818	0.692	0.574
Percentage loss in SW	5.29%	0.00%	5.95%	31.72%

column in Table 1.2 illustrate that charging the admission fee can reduce the arrival rate and eliminate the 5.39% social welfare loss in the baseline auction.

1.5.2 Service Provider's Revenue Maximization

Given the admission fee p and the baseline auction, the service provider's long-run average revenue is $p\lambda(p)$, where $\lambda(p)$ is the arrival rate under p. We show this structure raises the optimal revenue for the service provider under some technical assumptions we will introduce presently. Thus, finding the revenue-maximizing optimal mechanism reduces to pinning down the optimal admission fee.

We define *virtual type functions* and assume their monotonicity for the remainder of this chapter.

Definition 1.2. *Denote by $f_r(\cdot)$ and $f_p(\cdot, \tilde{c})$ the receivers' and payers' virtual type functions, respectively:*

$$f_r(c) \triangleq c + \frac{F(c)}{f(c)}, \qquad f_p(c; \tilde{c}) \triangleq c - \frac{F(\tilde{c}) - F(c)}{f(c)}.$$

Assumption 1.2. $\frac{df_r(c)}{dc} > 0$ *and* $\frac{df_p(c;\tilde{c})}{dc} > 0$ *for all* $c \in [\underline{c}, \tilde{c}]$ *and* $\tilde{c} \in [\underline{c}, \overline{c}]$.

Assuming monotone virtual type functions is common in the mechanism design literature. Afèche and Pavlin (2016) make the same assumption in studying priority pricing problems in queueing models whereas Myerson and Satterthwaite (1983) make this assumption in characterizing the intermediary's optimal mechanism for trading. Monotone virtual types are satisfied by many common probability distributions, such as the uniform, normal, logistic

and power function distributions, and the gamma and Weibull distributions with shape parameters greater than or equal to 1; any log-concave distribution has this property (Bagnoli and Bergstrom 2005).

In our framework, since the arrival rate is endogenous, the virtual type $f_p(\cdot, \tilde{c})$ is endogenously parametrized by $\tilde{c}$, deviating from standard virtual type assumptions in mechanism design that assert $\frac{df_r(c)}{dc} > 0$ and $\frac{df_p(c;\bar{c})}{dc} > 0$ for all $c \in [\underline{c}, \bar{c}]$. Lemma 1.1 shows that the standard assumptions turn out to be equivalent to Assumption 1.2.

Lemma 1.1. *$\frac{df_p(c;\tilde{c})}{dc} > 0$ for all $c \in [\underline{c}, \tilde{c}]$ and $\tilde{c} \in [\underline{c}, \bar{c}]$ is a necessary and sufficient condition for $\frac{df_p(c;\bar{c})}{dc} > 0$ for all $c \in [\underline{c}, \bar{c}]$.*

The service provider is not bound by the form of mechanism we introduce (a flat admission fee plus the baseline auction). For example, it could revise the auction rule so as not to induce strict priority. Proposition 1.2 indicates, nevertheless, that it is optimal under Assumption 1.2 for the service provider to appeal to the same mechanism structure as the social planner does. The only difference is that the service provider should set a higher admission fee.

Proposition 1.2. *The service provider maximizes revenue by setting a price $p^M > p^{SW}$ and running the baseline auction.*

Given the mechanism structure, if the service provider's only lever were the admission fee, then it should be intuitive that the service provider would set a higher fee than is socially optimal. Naor (1969) has a similar result in a different queueing context. As a monopolist, the service provider would command a higher price than the efficient level to maximize its own revenue. Proposition 1.2 reveals that even if the service provider has more levers, it should stick to mechanisms that implement strict priority. Its optimality is established from the optimal direct revelation mechanism (see Appendix A.1). By the revelation principle, for any mechanism that constitutes a Bayesian game, there exists an outcome-equivalent incentive-compatible direct revelation mechanism. Thus if an indirect mechanism matches the outcome of the optimal direct revelation mechanism, then the indirect mechanism is also

optimal. We shall see more on this when we analyze the intermediary's optimal mechanism in §1.6. The monotone virtual types in Assumption 1.2 are crucial here as they guarantee that the service provider has the same incentive as the social planner in prioritizing customers. Otherwise, the service provider would prefer pooling, i.e., serving a class of customers by the FIFO rule despite differences in their waiting costs (cf. Katta and Sethuraman 2005).

Note that as one of the many implementations of the service-provider's optimal direct mechanism, the proposed trading mechanism in Proposition 1.2 is outcome equivalent to a priority auction with an optimally determined reserve price (cf. Lui 1985 and Afèche and Mendelson 2004). Yet unlike the priority auction, there is no price discrimination by the service provider: all the payments generated in the baseline auction are transfers among customers; still, the same optimal revenue is achieved. We highlight that our proposed trading mechanism, albeit not the unique implementation of the optimal mechanism, is rather simple and that the flat admission fee is only for accessing the service facility, not for gaining priority, so it does not have the unfair connotation like the Priority Answer feature offered by EE (BBC News Business 2014).

Table 1.2 illustrates the service provider's optimal trading mechanism in column four, and the optimal pricing of a FIFO queue in column five. The FIFO price, p^F, is defined by

$$(p^F, \tilde{c}^F) = \arg\max_{(p,\tilde{c})} \quad p\Lambda F(\tilde{c}), \quad \text{s.t.} \;\; V - p - \tilde{c}W(\tilde{c}) = 0. \tag{1.4}$$

Denote the FIFO revenue by $\Pi^F = p^F \Lambda F(\tilde{c}^F)$. While it is immediate that the trading mechanism outperforms FIFO pricing in its revenue performance (2.381 vs. 2.011), it is not clear how the admission fees in the two scenarios, p^M and p^F, compare. Since the exclusive source of revenue in both scenarios is the admission fee, one might expect the service provider who shifts from FIFO to trading to increase this price to extract more revenue. This intuition is correct if the full market is already captured by FIFO pricing, i.e., $\tilde{c}^F = \overline{c}$, but in general, the direction of the service provider's price adjustment is ambiguous. Table 1.2 shows a

possibility that the service provider decreases the price (from 3.505 to 3.439) and achieves a higher revenue through a higher throughput. A lower price might be more palatable to customers and make them more receptive of the trading platform.

1.6 Trading through an Intermediary

The service provider may not want to operate the trading platform itself for technological reasons and reputational concerns, and would rather delegate it to a revenue-maximizing intermediary that develops and manages the platform. We conceptualize this in a sequential game where the service provider first sets an admission fee p (but does not dictates the trading mechanism), and then given this price, the intermediary designs a (fee-setting) trading mechanism that maximizes its own revenue.

The high-level distinction between the service provider and the intermediary is that the intermediary can only charge customers for using the trading platform (e.g., a trade participation fee), but not for access to the service facility (e.g., an admission fee). Since a high trade participation fee will make trading less attractive and eventually deter some customers from trading altogether, the intermediary's fee-structure will potentially affect customers' trading incentives.

In this section, we first study the intermediary's optimal trading mechanism given the service provider's admission fee, p. We start with the intermediary running a baseline auction coupled with an upfront *trade participation fee* H for customers who opt into trading. We discuss the shortcomings of this format, and propose an augmented auction in §1.6.1 that overcomes these shortcomings. We analyze the customer joining and trading behavior in §1.6.2. We establish the optimality of the augmented auction among all trading mechanism for the intermediary in §1.6.3. Finally, we solve in §1.6.4 the subgame perfect equilibrium where the service provider determines the admission fee, p, in the first stage in anticipation of the intermediary's and customers' best response in the second stage. For the ease of exposition, we normalize the admission fee to zero in solving the intermediary's second-stage

optimal mechanism, , i.e., we replace $V - p$ by V in §1.6.1-1.6.3. In §6.4, we then replace V by $V - p$ and characterize the service provider's optimal choice of p.

Consider a benchmark trading mechanism where an arriving customer must pay the intermediary an upfront trade participation fee H; then, the baseline auction is run as before. We refer to this as an "H auction." The intermediary's revenue is $H\lambda^T(H)$, where $\lambda^T(H)$ is the arrival rate of the customers who trade given H. By definition, trading customers are a subset of joining customers, i.e., $\lambda^T(H) \leq \lambda$ for any H.

Recall that in the baseline auction (where $H = 0$), all joining customers are strictly better off by participating in trading. Thus, the equilibrium structure identified in Proposition 1.1 remains valid if H is slightly positive. It is easy to see that if the trade participation fee H is too high, then trading will no longer be favored over FIFO. Hence, there exists a threshold value $\overline{H}$ such that the equilibrium structure identified in Theorem 1.1 is preserved and all joining customers voluntarily trade ($\lambda^T(H) = \lambda$) if and only if $H \leq \overline{H}$.

Definition 1.3. *$\overline{H}$ is such that $\lambda^T(H) = \lambda$ if and only if $H \leq \overline{H}$.*

For convenience, we refer to the auction where $H = \overline{H}$ as the "$\overline{H}$ auction."

Continuing the numerical example, at $H = \overline{H} = 1.342$, $U(c)$ in Figure 1.1-(c) would be tangent to the FIFO line, and the intermediary's revenue is 1.208. If $H > \overline{H}$, some customers with medium waiting costs (since they benefit the least from trading) will find trading too costly and thus refuse to trade by submitting "No," and this would lead to $\lambda^T(H) < \lambda$. A complete equilibrium analysis when $H > \overline{H}$ is available in Appendix A.6. The revenue-maximizing intermediary's is in a conundrum. On one hand, if it would like to get all joining customers to trade, its fee is bounded above by $\overline{H}$. One the other hand, if the intermediary wants to charge more aggressively (above $\overline{H}$), it must bear the cost of being unable to collect the fee from some joining customers: a direct loss of revenues via a decreased trading volume, plus, an indirect loss via a lower arrival rate as the non-trading (FIFO) customers downgrade the expected utility of those who trade.

Now we enrich the baseline auction with two trade-restriction prices that enable the

intermediary to charge above $\overline{H}$ while still inducing voluntary trading of all joining customers. We shall show this is the optimal trading mechanism for the intermediary.

1.6.1 Augmented Auction: Trading Rules and a Motivating Example

Auction format: The augmented auction contains two trade restriction parameters $\underline{R}$ and $\overline{R}$ ($\underline{R} \leq \overline{R}$) in addition to the trade participation fee H. The trading rule is the same as before except that *if both customers' bids are within the interval* $[\underline{R}, \overline{R}]$*, they are barred from trading with one another and are served FIFO.* However, if only one of the two customers' bids are within $[\underline{R}, \overline{R}]$, trade still occurs between the two. This auction is referred to as an "$(H, \underline{R}, \overline{R})$ auction."

Illustration 2: Consider the example of the previous section and assume that $b_4 < \underline{R} < b_2 < b' \leq b_1 < \overline{R}$. As before, the system prior to trading is represented by (b_1, b_2, F, b_4, b'). Only customer 4 and 5 swap positions, and the system after trading is represented by (b_1, b_2, F, b', b_4). Note that despite the fact that $b_2 < b'$, customers 2 and 5 do *not* swap positions since both of their bids fall in $[\underline{R}, \overline{R}]$.

Table 1.3: The intermediary's optimal $(H, \underline{R}, \overline{R})$ auction and the $\overline{H}$ auction.

	Revenue	H	$\underline{R}$	$\overline{R}$	λ
Optimal augmented auction	1.352	1.510	0.257	0.425	0.896
$\overline{H}$ auction	1.208	1.342	-	-	0.9

Table 1.3 shows that when $H = 1.510$, $\underline{R} = 0.257$ and $\overline{R} = 0.425$, the intermediary's revenue would be 1.352 in the $(H, \underline{R}, \overline{R})$ auction, 11.9% higher than the revenue that would be achieved in the $\overline{H}$ auction. Note that the trade participation fee H in the augmented auction is higher than $\overline{H}$, yet all joining customers sign up for trading, which can be verified by recognizing the revenue (1.352) is equal to H (1.510) times λ (0.896).

1.6.2 Auction Equilibrium

To generate insights into how the augmented auction with trade restriction benefits the intermediary, we derive the equilibrium for the case when trading is free ($H = 0$, budget-balanced among customers), and compare that with the budget-balanced baseline auction. We indicate the equilibrium in the augmented auction by means of a superscript A. With a slight abuse of notation, we use $U(c, \beta)$ to denote the expected utility of customer c who bids β in the equilibrium of the augmented auction (including the trade participation fee).

Theorem 1.2 (Full Trading, Partial Pooling Equilibrium)**.** *Under the augmented auction with given $\underline{R}$ and $\overline{R}$, when $H = 0$, there exists an equilibrium in which:*
(i) $J^A(c)$ =join for $c \in [\underline{c}, \tilde{c})$ (and balk otherwise);
(ii) the equilibrium bid function is weakly increasing and given by

$$b^A(c; c_r, c_p, \tilde{c}) = \begin{cases} c + \frac{\int_c^{c_r} (F(\tilde{c})-F(s))^2 W^e(s;\tilde{c}) ds + K(\underline{R}, c_r, c_p, \tilde{c})}{(F(\tilde{c})-F(c))^2 W^e(c;\tilde{c})}, & c \in [\underline{c}, c_r) \\ \overline{R}, & c \in [c_r, c_p] \\ b^B(c; \tilde{c}), & c \in (c_p, \tilde{c}] \end{cases} \tag{1.5}$$

where constant $K(\underline{R}, c_r, c_p, \tilde{c}) = (\underline{R} - c_r)(F(\tilde{c}) - F(c_r))^2 W^e(c_r; \tilde{c})$ and $c_r, c_p, \tilde{c} \in \Xi$ with $\underline{c} \le c_r \le c_p \le \tilde{c}$ are a solution to the following equations:

$$[U(c_r, \underline{R}) - U(c_r, \overline{R})][c_r - \underline{c}][c_r - \tilde{c}] = 0 \tag{1.6a}$$

$$[U(c_p, b^A(c_p^+; c_r, c_p, \tilde{c})) - U(c_p, \overline{R})][c_p - \underline{c}][c_p - \tilde{c}] = 0 \tag{1.6b}$$

$$U(\tilde{c}, b^A(\tilde{c}; c_r, c_p, \tilde{c}))[\tilde{c} - \overline{c}] = 0; \tag{1.6c}$$

(iii) the expected waiting time for customer $c \in [\underline{c}, \tilde{c}]$ is

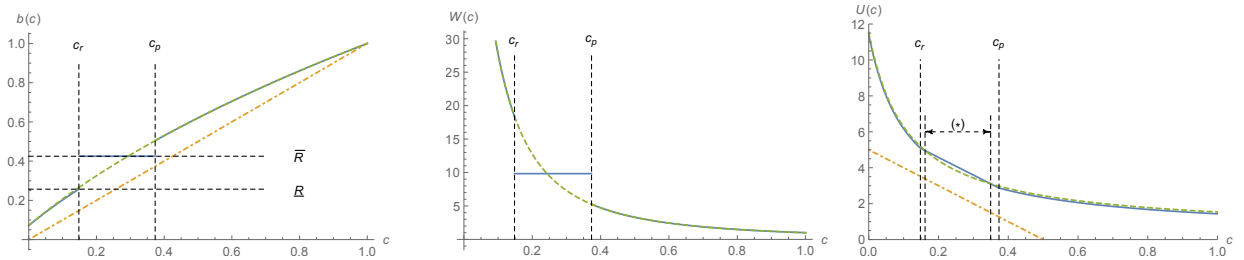
$$W^A(c; c_r, c_p, \tilde{c}) = \begin{cases} W^e(c; \tilde{c}), & \forall\ c \in [\underline{c}, c_r) \cup (c_p, \tilde{c}] \\ \frac{1}{\mu(1-\rho F(\tilde{c})+\rho F(c_r))(1-\rho F(\tilde{c})+\rho F(c_p))}, & \forall c \in [c_r, c_p]. \end{cases} \tag{1.7}$$

Comparing Theorem 1.2 with Theorem 1.1 shows the effects of the trade restriction parameters, $\underline{R}$ and $\overline{R}$. While the bid function $b^A(\cdot)$ in (1.5) is still strictly increasing in $[\underline{c}, c_r] \cup (c_p, \tilde{c}]$, it is flat in $[c_r, c_p]$: these customers all bid $\overline{R}$ and thus do not trade with one another (see Figure 1.2-(a)). As a result, the waiting time schedule is no longer efficient since these customers all expect the same waiting time despite their different waiting costs (see Figure 1.2-(b)). Consequently, there is pooling of customers in $[c_r, c_p]$, who are served as a single FIFO class. As shown in the expression of $W^A(\cdot)$ in (1.7), for any given arrival rate, the expected waiting time for customers in $[\underline{c}, c_r) \cup (c_p, \tilde{c}]$ is still the efficient waiting time, i.e., trading allows these customers to be strictly prioritized over any other joining customer with lower waiting cost.

Three features are noteworthy about customers' bidding behaviors. First, if any customer wishes to bid in between $[\underline{R}, \overline{R}]$, the rational decision is to bid up to $\overline{R}$. By the auction rule, a customer could bid strictly between $\underline{R}$ and $\overline{R}$ and expect the same waiting time (since she is barred from trading with customers bidding in $[\underline{R}, \overline{R}]$ regardless of her bid). Such a bid would adversely affect the money she receives when selling her spots to anyone who bids higher as the trading price is dictated solely by the seller's bid. Therefore, customers always have the incentive to bid at the boundary $\overline{R}$ if they ever want to submit a bid that falls within $[\underline{R}, \overline{R}]$. Second, the left limit of the bid function at c_r is $\underline{R}$, i.e., $b^A(c_r^-) = \underline{R}$. If customer c_r^- bids strictly lower than $\underline{R}$, then she can always up her bid to $\underline{R}$ without changing her priority yet strictly increasing her revenue as a seller. Likewise, this is driven by the auction rule that requires the trading price be equal to the seller's bid. Third, augmenting the auction with $\underline{R}$ and $\overline{R}$ does not change the bidding behavior of customers in $[c_p, \tilde{c}]$ for a given arrival rate. If a customer decides to bid above $\overline{R}$, then he will purchase from all customers who bid below, regardless of whether the other customer bids within $[\underline{R}, \overline{R}]$, and therefore, this collapses to the same environment as in the baseline auction, making the customer follow the original bidding strategy. By comparison, the augmented auction does affect the bidding behavior of customers in $[\underline{c}, c_r]$. As customer c_r^- bids $\underline{R}$, other customers' bidding strategy are adjusted

accordingly if they decide to bid below $\underline{R}$. However, as Figure 1.2-(b) illustrates, the bid function for $c \in [\underline{c}, c_r)$ in the augmented auction only marginally differs from its counterpart in the baseline auction in the numerical example.

Figure 1.2: The augmented auction.



(a) Bidding function. (b) Waiting time. (c) Utility.

Note. $H = 0$, $\underline{R} = 0.257$, $\overline{R} = 0.425$. $\tilde{c} = 1$, $c_r = 0.148$, $c_p = 0.373$. The blue solid curves correspond to the equilibrium properties of the augmented auction; the green dashed curve, the baseline auction as in Figure 1.1. The orange dot-dashed line is the 45-degree line in (a); the expected utility if customers are served FIFO under the same equilibrium arrival rate in (c). $(*)$ indicates the subset of customer types in the pooling segment that receive a higher expected utility than in the baseline auction.

We use (1.6a)-(1.6c) to determine $c_r, c_p, \tilde{c}$. (1.6a) implies that customer c_r must be indifferent between bidding $\underline{R}$ and $\overline{R}$ subject to $c_r > \underline{c}$. Likewise, (1.6b) captures the indifference of customer c_p between bidding $\overline{R}$ and $b^A(c_p^+)$ subject to $c_p < \tilde{c}$. (1.6c) captures the indifference of customer $\tilde{c}$ between joining and balking provided that not all customers join ($\tilde{c} < \overline{c}$). $\underline{R}$ and $\overline{R}$ should be set at some intermediate values in order for c_r and c_p to take interior solutions.

Similar to the baseline auction, adding $\underline{R}$ and $\overline{R}$ to the budget balanced auction does not discourage any joining customers from voluntarily participating in trading. As shown in Figure 1.2-(c), all joining customers are better off by participating in trading than submitting "No." One noticeable difference of customers' expected utility in the augmented auction is that it decreases linearly in c for $c \in [c_r, c_p]$. This is by the linearity assumption of waiting costs. Customers in the pooling segment $[c_r, c_p]$ differ in their waiting costs but choose to bid the same amount and thus expect to wait the same amount of time and pay/receive the same amount of money.

In Figure 1.2-(c), there is a subset of customers, indicated by $(*)$, in the pooling segment who receive higher expected utility in the augmented auction than in the baseline auction. These customers benefit from the augmented auction for two reasons. First, some of them bid higher than they do originally (in Figure 1.2-(a), the new bid $\overline{R}$ is higher than the original bid $b^B(c)$ for some medium customers in the pooling segment), and by the auction rule, this implies that they can collect a higher proceed when they sell their waiting spots. Second, recall from §1.4 that in the baseline auction, a buyer may lose in some trades in which the seller has a slightly lower waiting cost and submits a bid that exceeds the buyer's true waiting cost (though lower than the buyer's bid), i.e., $b^B(c_b) > b^B(c_s) > c_b > c_s$. In the augmented auction, disallowing these customers from being engaged in those unprofitable trades improves their utility. The presence of the customer segment $(*)$ intuitively sheds light on why, for the intermediary, the augmented auction could potentially beat the baseline auction: the middle customers derive more value from the augmented auction and thus are willing to pay more for the opportunity to trade. On the flip side, observe from 1.2-(c) that customers with high waiting costs are worse-off in the augmented auction. Hence, more customers might balk. In light of this price-quantity trade-off, it becomes subtle whether the intermediary always wants to implement the augmented auction.

1.6.3 Optimal Auction Parameters and Structure

When $H = 0$ in the augmented auction (the intermediary has no revenue), all customers strictly prefer trading. A slightly positive H will not alter this preference and the equilibrium bidding behavior in the augmented auction insofar as all joining customers trade. We show that this is indeed the optimal structure the intermediary wants to implement. Furthermore, the $(H, \underline{R}, \overline{R})$ auction is an optimal mechanism for the intermediary given the optimal auction parameters.

Theorem 1.3 (Optimality of the Augmented Auction)**.** *The* $(H^*, \underline{R}^*, \overline{R}^*)$ *is an optimal*

mechanism for the intermediary with:

$$H^* = \frac{\Pi(c_r^*, c_p^*, \tilde{c}^*)}{\Lambda F(\tilde{c}^*)} \tag{1.8a}$$

$$\overline{R}^* = \frac{c_p^* \overline{W}(\tilde{c}^*) + [\rho b^R(c_p^*; c_r^*, c_p^*, \tilde{c}^*)(F(\tilde{c}^*) - F(c_p^*)) - c_p^*] W^e(c_p^*; \tilde{c}^*)}{\rho(F(\tilde{c}^*) - F(c_r^*))\overline{W}(\tilde{c}^*)} \tag{1.8b}$$

$$\underline{R}^* = \frac{c_r^* W^e(c_r^*; \tilde{c}^*) - c_r^* \overline{W}(\tilde{c}^*) + \rho \overline{W}(\tilde{c}^*)[F(\tilde{c}^*) - F(c_p^*)]\overline{R}^*}{\rho W^e(c_r^*; \tilde{c}^*)[F(\tilde{c}^*) - F(c_r^*)]} \tag{1.8c}$$

where $\Pi(c_r, c_p, \tilde{c}) = -\Lambda \int_{\underline{c}}^{c_r} \left(W^e(c; \tilde{c}) - \overline{W}(\tilde{c})\right) f_r(c) dF(c) + \Lambda \int_{c_p}^{\tilde{c}} \left(\overline{W}(\tilde{c}) - W^e(c; \tilde{c})\right) f_p(c; \tilde{c}) dF(c)$ *and* $c_r^*, c_p^*, \tilde{c}^*$ *solve the following optimization problem:*

$$\max_{c_r, c_p, \tilde{c} \in \Xi : c_r \le c_p \le \tilde{c}} \quad \Pi(c_r, c_p, \tilde{c}) \tag{1.9a}$$

$$\text{s.t.} \quad \frac{1}{\mu(1 - \rho F(\tilde{c}) + \rho F(c_r))(1 - \rho F(\tilde{c}) + \rho F(c_p))} = \overline{W}(\tilde{c}) \tag{1.9b}$$

$$V - \int_{c_p}^{\tilde{c}} W^e(c; \tilde{c}) dc - c_p \overline{W}(\tilde{c}) \ge 0. \tag{1.9c}$$

The resulting equilibrium structure is the same as identified in Theorem 1.2. In particular,

$$U(c, b^A(c; c_r^*, c_p^*, \tilde{c}^*)) = V - c\overline{W}(\tilde{c}^*), \quad \forall c \in [c_r^*, c_p^*]. \tag{1.10}$$

Theorem 1.3 determines the optimal parameters $(H^*, \underline{R}^*, \overline{R}^*)$ by reverse engineering. Instead of characterizing the equilibrium outcome by solving (1.6a)-(1.6c) for $c_r, c_p, \tilde{c}$ under any given auction parameters $(H, \underline{R}, \overline{R})$, we first determine what the optimal outcome should be by obtaining $c_r^*, c_p^*, \tilde{c}^*$ from the optimization problem (1.9a)-(1.9c) and then determine the optimal auction parameters $(H^*, \underline{R}^*, \overline{R}^*)$ that can implement the optimal outcome using (1.8a)-(1.8c), where (1.8b) and (1.8c) are obtained from shuffling terms of (1.6a) and (1.6b).

The optimization problem (1.9a)-(1.9c) arises from the analysis of the optimal direct revelation mechanism (detailed in Appendix A.1). In addition to the usual incentive compatibility and individual rationality constraints, our direct mechanism formulation also incorporates

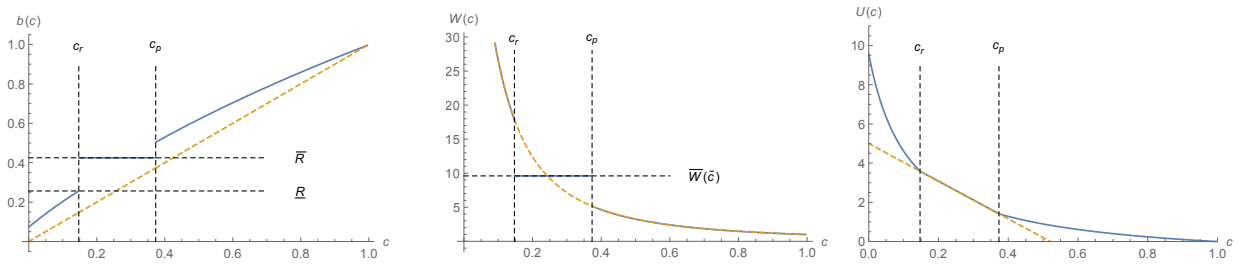
operational constraints using the achievable region method (Coffman and Mitrani 1980), as well as property right constraints ensuring that all joining customers should be no worse off than being served FIFO. Some manipulation simplifies the direct mechanism formulation to (1.9a)-(1.9c). Constraint (1.9c) specifies that joining must be individually rational. Next, we provide more intuition about (1.9a) and (1.9b).

Combining (1.9b) and $W^A(\cdot)$ in (1.7) implies that in the optimal auction, the expected waiting time for customers in $[c_r^*, c_p^*]$ is equal to the FIFO waiting time $\overline{W}(\tilde{c})$ (see Figure 1.3-(b)). Furthermore, (1.10) suggests that these customers' expected utility is equal to what they would get if they just bid "No" and wait FIFO (see Figure 1.3-(c)). This does not imply they do not trade at all: they still swap positions with customers *outside the pool* by selling their spot to higher bidders and buying positions from lower bidders, but on average trading does not realize any gains. The fact that they trade is crucial to achieving an efficient expected waiting time for customers outside the pool $[\underline{c}, c_r^*) \cup (c_p^*, \tilde{c}^*]$. The augmented auction is designed to only prohibit trading within the pool.

In (1.9a), the objective function, $\Pi(c_r, c_p, \tilde{c})$, expresses the intermediary's long-run average revenue as a function $c_r, c_p, \tilde{c}$. Following the interpretations in Bulow and Roberts (1989), $\Pi(c_r, c_p, \tilde{c})$ can be intuitively understood as follows. Suppose that buyers make payments to the intermediary who then redistributes the money back to sellers and keeps the trade participation fee to itself. On average, customers in $[\underline{c}, c_r)$ wait longer than if they are served FIFO (the difference is $W^e(c; \tilde{c}) - \overline{W}(\tilde{c})$) and expect to receive more money than they pay. On the contrary, customers in $(c_p, \tilde{c}]$ have shorter waiting time on average (the expected waiting time reduced is $\overline{W}(\tilde{c}) - W^e(c; \tilde{c})$) and expect to pay more than they receive. Therefore, the first integral in $\Pi(c_r, c_p, \tilde{c})$ is the intermediary's long-run average losses from compensating patient customers, whereas the second integral is the long-run average gains from charging impatient customers. The net inflow is the intermediary's revenue. Note that virtual types in $\Pi(c_r, c_p, \tilde{c})$ captures the customers' information rents that arise because only an individual customer knows her private type. As a result, the intermediary has to pay

more than the actual waiting cost c when it compensates customers for longer waiting time, i.e., $f_r(c) > c$, but can charge less than c for a shorter waiting time, i.e., $f_p(c;\tilde{c}) < c$. Now that we understand $\Pi(c_r, c_p, \tilde{c})$, (1.8a) is straightforward: the optimal trading participation fee H^* is simply the intermediary's optimal revenue divided by the optimal arrival rate. The intermediary can collect H^* from all arriving customers since they all trade.

Figure 1.3: The optimal $(H, \underline{R}, \overline{R})$ auction.



(a) Bidding function. (b) Waiting time. (c) Utility.

Note. $H = 1.510$, $\underline{R} = 0.257$, $\overline{R} = 0.425$. $\tilde{c} = 0.995$, $c_r = 0.147$, $c_p = 0.373$. The solid curve corresponds to the equilibrium properties of the auction; the dashed curve is the 45-degree line in (a); the efficient waiting time function in (b); the expected utility if customers are served FIFO under the same equilibrium arrival rate in (c).

In brief, the optimal auction has the following two features:

(1) All joining customers participate in trading.

(2) The schedule is efficient in $[\underline{c}, c_r^*) \cup (c_p^*, \tilde{c}^*]$ but customers in $[c_r^*, c_p^*]$ expect FIFO waiting time.

Notice that should $c_r^* = c_p^*$ be optimal to (1.9a)-(1.9c), then $W^e(c_r^*; \tilde{c}^*) = \overline{W}(\tilde{c}^*)$ from (1.9b) and this would imply $\underline{R}^* = \overline{R}^*$ in (1.8c), which would render the $(H, \underline{R}, \overline{R})$ auction degenerate as trade would occur between any two customers with different waiting costs. Thus, the outcome would be an efficient schedule and could be implemented simply with an H auction with $H = \overline{H}$. Theorem 1.4, however, rules out this possibility.

Theorem 1.4 (Optimality of Trade Restriction)**.** *The optimal auction should always have* $\underline{R}^* < \overline{R}^*$*, and thus* $c_r^* < c_p^*$.

Theorem 1.4 shows that the intermediary would like to restrict trading to a certain extent to fully exploit its control over the trading channel. As a result, it sets $\overline{R}^*$ strictly above $\underline{R}^*$ so that pooling occurs in the intermediary's optimal auction and the schedule is not efficient. This extends the classical result in Myerson and Satterthwaite (1983) about the intermediary's trade restriction incentives in bilateral trading to a queueing context where customers can be both buyers and sellers. After all, the intermediary is a monopolist who wants to restrict output under the efficient level to command a higher price. Note that if the service provider operates the trading platform as in §1.5.2, it can simply exercise its monopoly power by charging a higher admission fee (which reduces arrivals). By contrast, the intermediary's trade participation fee cannot be forced upon customers even after they join the system, so setting a higher fee as an intermediary should be done in a more nuanced way that reduces the amount of trading among customers and creates a pooling segment. As we argue in §1.6.2, trade restriction generates value to customers in the middle who are most sensitive to trading by letting them avoid undesirable trades. This value added, in turn, passes on to the intermediary by enabling it to charge a higher trade participation fee (we shall formally establish this in Corollary 1.1).

Recall the observation from §1.6.2 that trade restriction hurts customers with high waiting costs who are most sensitive to joining. This adds an additional layer of complexity to the question whether trade restriction always benefits the intermediary. Considering the endogenous arrival rate is unique to our queueing context and not discussed in the classical trading literature. Theorem 1.4 indicates that trade restriction (or partial pooling) is always optimal even when the arrival rate is endogenous. Unlike Katta and Sethuraman (2005), this pooling result does not require the virtual type functions to be irregular (violating Assumption 1.2); unlike Afèche and Pavlin (2016), it does not rest on the assumption that customer types are lead-time dependent. Moreover, the pooling result does not arise only as a special case, but instead holds for all model parameters considered.

Let $\lambda^* = \Lambda F(\tilde{c}^*)$ be the optimal effective arrival rate under the $(H^*, \underline{R}^*, \overline{R}^*)$ auction.

Let λ^{FIFO} be the effective arrival rate if all joining customers are served FIFO. In the FIFO system, a customer of type c receives an expected utility of $V - c\overline{W}(\tilde{c}^{\text{FIFO}})$, where $\tilde{c}^{\text{FIFO}}$ satisfies $\lambda^{\text{FIFO}} = \Lambda F(\tilde{c}^{\text{FIFO}})$. Let $\lambda^{\overline{H}}$ be the effective arrival rate if the intermediary charges $\overline{H}$ in the H auction. Likewise, $\tilde{c}^{\overline{H}}$ is defined such that $\lambda^{\overline{H}} = \Lambda F(\tilde{c}^{\overline{H}})$. Proposition 1.3 orders the three arrival rates.

Proposition 1.3. $\lambda^{FIFO} \leq \lambda^* \leq \lambda^{\overline{H}}$. *In particular, if not all customers join in the optimal auction, i.e.,* $\lambda^* < \Lambda$, *then* $\lambda^{FIFO} < \lambda^* < \lambda^{\overline{H}}$.

In the optimal augmented auction, customers with high waiting costs who would otherwise balk in a FIFO system may now join and participate in trading because this option to trade makes them better off than if they are served FIFO. On the other hand, as compared to the $\overline{H}$ auction, the pooling segment in the optimal mechanism diminishes the appeal of trading for customers with high waiting costs since they have to pay more to get the same priority; thus fewer customers join. The optimal mechanism has a lower arrival rate than the $\overline{H}$ auction, but it raises more revenue, which must stem from a higher trade participation fee. This is summarized in Corollary 1.1.

Corollary 1.1. *The optimal trade participation fee in the augmented auction is strictly above* $\overline{H}$, *i.e.,* $H^* > \overline{H}$.

Trade restriction resolves the intermediary's conundrum: it always charges strictly above $\overline{H}$, yet all joining customers choose to pay this trade participation fee. Combining Theorem 1.4, Proposition 1.3 and Corollary 1.1 shows that the gain from charging a higher trade participation fee above $\overline{H}$ in the augmented auction overshadows the cost of a resulting lower arrival rate for the intermediary.

1.6.4 The Value of Trading vs. FIFO

So far, we have substituted V for $V - p$. Now we reintroduce p and turn on the service provider's pricing decision in the presence of the intermediary. We assume that originally

the service provider sets a revenue-maximizing admission fee, p^F, in a FIFO queue, as formalized in (1.4). We compare this to a setting where the service provider invites the revenue-maximizing intermediary to mediate the trading platform. The service provider is completely detached from the trading platform and does not contract with the intermediary due to technological reasons and reputational concerns. The service provider and the intermediary play a sequential game: the service provider sets an admission fee p^T in anticipation of the intermediary's optimal trading mechanism. Our main question is whether, with the revenue maximizing intermediary, the service provider should increase/decrease its admission fee and whether its revenue will ever be adversely affected by the intermediary. Formally, the subgame perfect equilibrium as follows:

$$\big(H(p), \underline{R}(p), \overline{R}(p)\big) = \arg\max_{H,\underline{R},\overline{R}} \; H\lambda(p, H, \underline{R}, \overline{R}) \tag{1.11a}$$

$$p^T = \arg\max_{p} \; p\lambda(p, H(p), \underline{R}(p), \overline{R}(p)) \tag{1.11b}$$

where $\lambda(p, H, \underline{R}, \overline{R})$ is the arrival rate if the service provider charges p and the intermediary runs an $(H, \underline{R}, \overline{R})$ auction; $\big(H(p), \underline{R}(p), \overline{R}(p)\big)$ are the intermediary's optimal auction parameters given the service provider's price p. (1.11a) states that the intermediary sets the optimal auction parameters that maximize its revenue given the service provider's admission fee. (1.11b) states that the service provider sets the optimal admission fee to maximize its revenue. Let Π^T denote the service provider' optimal revenue in (1.11b).

We emphasize that for the service provider, the key distinction between this sequential game and the setup in §1.5.2 is that here the service provider *only* sets the admission fee p^T and leave the trading mechanism to the intermediary, whereas in §1.5.2, the service provider determines *both* admission fee and the auction rule for trading. Solving for the optimal p^T is nontrivial in this problem since for every equilibrium conjecture p, the optimization problem (1.9a)-(1.9c) has to be solved with V replaced by $V - p$ in (1.9c), which gives the arrival rate and subsequently revenue for the service provider who sets price to p. Yet, we obtain

the following result:

Corollary 1.2. *The intermediary running the trading platform strictly increases the service provider's revenue relative to FIFO, i.e.,* $\Pi^T > \Pi^F$.

Corollary 1.2 follows from Proposition 1.3. If the service provider continues to charge p^F when the intermediary runs the trading platform, we know from Proposition 1.3 that the arrival rate weakly increases. If the arrival rate strictly increases, then the service provider will earn more revenue even without any price adjustment, so setting the price p^T optimally can only further boost its revenue (we do not know yet how p^T and p^F compare in this case). It remains to show that if the arrival rate stays the same after trading is introduced, which only happens when the full market is captured, the service provider is able to exploit the value of trading by charging a higher price, i.e., $p^T > p^F$, and consequently earn more revenue. Once again, this is guaranteed by the trading mechanism. Under FIFO pricing, the marginal customer $\bar{c}$ must expect zero utility: $V - \bar{c}\overline{W}(\bar{c}) - p^F = 0$. If the same price p^F is charged under the intermediary's optimal trading mechanism, this customer strictly prefers trading to FIFO, i.e., $U(\bar{c}) - p_F > V - \bar{c}\overline{W}(\bar{c}) - p^F$, where $U(\bar{c})$ is the expected utility from trading excluding the admission fee p^F. This implies $U(\bar{c}) - p^F > 0$, and the service provider can increase p^F to p^T such that this customer again becomes indifferent between joining and balking, i.e., $U(\bar{c}) - p^T = 0$.

We perform a line search of p to determine the optimal p^T numerically. We quantify, under a variety of input parameters, how the intermediary's trading mechanism impacts the service provider's pricing decision in Table 1.4. We also compare the service provider's revenue with the intermediary's trading platform relative to a FIFO queue in Table 1.5. In all numerical trials, we fix the service rate to be 1 and the waiting cost to be uniformly distributed between 0 an 1 as in Table 1.1. We vary the service value V between 2 and 10, and the market size Λ between 0.1 and 3. Denote the percentage difference between the service provider's FIFO price, p^F, and price under trading, p^T, by Δ_p; the percentage difference between the service provider's revenue under FIFO, Π^F and revenue under trading,

Π^T, by Δ_Π.

$$\Delta_p = \frac{p^T - p^F}{p^F} \times 100\%, \qquad \Delta_\Pi = \frac{\Pi^T - \Pi^F}{\Pi^F} \times 100\%.$$

Table 1.4: Percentage change in price Δ_p.

Λ	$V=2$	$V=3$	$V=4$	$V=5$	$V=6$	$V=7$	$V=8$	$V=9$	$V=10$
0.1	-0.73%	1.16%	0.76%	0.56%	0.45%	0.37%	0.32%	0.28%	0.25%
0.3	-1.49%	-1.73%	3.70%	2.66%	2.08%	1.71%	1.45%	1.26%	1.11%
0.5	-1.78%	-1.86%	-1.81%	-0.09%	6.39%	5.11%	4.26%	3.65%	3.19%
0.7	-1.86%	-1.78%	-1.60%	-1.42%	-1.19%	-0.98%	-0.82%	-0.59%	-0.09%
0.9	-1.83%	-1.63%	-1.39%	-1.01%	-0.87%	-0.57%	-0.50%	-0.29%	-0.23%
1.1	-1.76%	-1.68%	-1.31%	-0.87%	-0.52%	-0.34%	-0.18%	0.03%	0.13%
2	-1.59%	-0.73%	-0.32%	0.00%	0.38%	0.48%	0.65%	0.81%	0.87%
3	-0.79%	-0.15%	0.27%	0.50%	0.81%	0.87%	1.10%	1.20%	1.21%

Table 1.5: Percentage change in revenue Δ_Π.

Λ	$V=2$	$V=3$	$V=4$	$V=5$	$V=6$	$V=7$	$V=8$	$V=9$	$V=10$
0.1	1.65%	1.16%	0.76%	0.56%	0.45%	0.37%	0.32%	0.28%	0.25%
0.3	3.99%	5.21%	3.70%	2.66%	2.08%	1.71%	1.45%	1.26%	1.11%
0.5	5.55%	6.89%	7.82%	7.32%	6.39%	5.11%	4.26%	3.65%	3.19%
0.7	6.67%	7.97%	8.80%	9.36%	9.74%	10.01%	10.20%	10.33%	8.18%
0.9	7.49%	8.71%	9.42%	9.87%	10.15%	10.33%	10.45%	10.51%	10.55%
1.1	8.11%	9.22%	9.83%	10.18%	10.38%	10.49%	10.54%	10.56%	10.55%
2	9.64%	10.28%	10.51%	10.56%	10.53%	10.45%	10.35%	10.24%	10.13%
3	10.28%	10.55%	10.53%	10.40%	10.24%	10.07%	9.90%	9.73%	9.56%

Table 1.4 shows that it is in general ambiguous how the service provider should adjust its price when the intermediary implements the trading platform. Table 1.5 shows the service provider's revenue improvement. In terms of its pricing behavior, the numerical instances can be divided into three cases. *Case 1*: In the top right corner of Table 1.4, when the service value is high and the market size is small, the service provider's price goes up. This corresponds to the case when the full market is captured in the FIFO queue. As we argued following Corollary 1.2, trading allows the service provider to raise its price. This can be verified by recognizing that in those instances the relative price change is equal to the relative revenue change shown in Table 1.5 as the arrival rate is unaffected. *Case 2*: In the bottom right corner of Table 1.4, when the service value is high and the market size is also large, the

service provider's price rises again. However, in these instances, the system does not capture the full market, and the arrival rate is also changed as a result of trading. We observe that the revenue change is higher than the price change now, which implies that trading allows the service provider to both command a higher price and lure more customers. *Case 3*: In the rest of the instances, the service provider offers a price cut, so that the revenue increase is solely attributed to a higher arrival rate. Here the FIFO queue does not capture the full market. As Proposition 1.3 suggests, even if the service provider sticks to its original price, it will enjoy a higher revenue since more customers join when the trading platform is in place. However, the service provider responds by actually decreasing its price to get an even higher arrival rate. As we suggest in §1.5.2, a lower price might be more favorable to customers from a behavioral perspective, and this may facilitate the promotion of the intermediary's trading platform.

Somewhat strikingly, in our numerical study, when the market is not fully captured (cases 2 and 3), the magnitude of the price change is, in fact, quite small: less than 2% in all those instances; yet the revenue change is quite sizable by comparison (about 10% in many instances). The implication is that the intermediary's trading platform can potentially be a seamless built-in for the service provider: the service provider does not need to worry about running the auction itself; it does not even need to significantly alter its price as a response of the new platform (which is valuable especially when the menu cost is high). The bottom line is that the intermediary increases the service provider's revenue relative to a FIFO system, and improves system efficiency. These are the intermediary's value propositions to the service provider with either revenue or welfare considerations. Of course, there is a natural double marginalization problem in our setup where both the service provider and the intermediary are monopolists. Theoretically, the service provider would earn an even higher revenue if it operated the trading platform by itself as in §1.5.2. Practically, this may not be in the service provider's best interest for technological reasons and reputational concerns previously stated. Vertical integration would achieve the maximum joint revenue as in §1.5.2, but this usually

involves efforts expended on negotiation, coordination and contracting. In this regard, an intermediary on a separate platform should probably be good enough in practice.

1.7 Conclusion and Discussion

We analyze a congested service system in which customers are privately informed about their waiting cost and trade their waiting positions on a trading platform. We design the optimal mechanisms that maximize social welfare, the service provider's revenue, and the revenue of the intermediary that develops and manages the trading platform, respectively. We find that while both the social planner and the service provider want customers to trade as much as possible (inducing the $c\mu$ rule), the intermediary *restricts* trading among customers (pooling) to maximize its own revenue. We find that our budget-balanced baseline auction leads to a higher arrival rate than is socially desirable and thus an admission fee must be levied to maximize social welfare. By comparison, the revenue-maximizing service provider would charge a higher admission fee than the social planner would. For practical reasons, the service provider may wish to delegate the trading platform to a revenue-maximizing intermediary. To that end, we propose an augmented auction with a trade participation fee *and* two trade restriction prices. Compared to the baseline auction with a trade participation fee only, the intermediary can charge a higher fee in the optimal auction and still have *all* joining customers voluntarily participate in trading. We show that the intermediary's trading mechanism always strictly improves the service provider's revenue relative to a FIFO system despite the intermediary's revenue-maximizing nature. This is a potentially powerful sales argument the intermediary can make to convince the service provider of installing the platform.

Several key *modeling choices* of this chapter warrant further discussion.

Queue observability. We build a model of an unobservable queue mainly for technical convenience. It would be of interest to study a system where customers observe the queue length when they submit bids. In this vein, an auction can be run at each state transition

epoch (new arrival or departure). The challenge with modeling this system is that an arriving customer can infer the types of other customers from the queue length and/or the entire trading history, which affects bidding behavior. Computing the steady state probabilities for even a two-class priority queue is quite cumbersome (see e.g., Miller 1981), making customers' inference and dynamic bidding problems messy, if not intractable. Apart from this modeling challenge, badgering customers into frequently updating their bids may also fall short in practical implementation. In this regard, automating the trading process as we propose in this chapter would be more advisable.

Customer Valuation: In practice, customers may differ not only in waiting costs, but also their service valuation. For two individuals with the same waiting cost, the one with a higher valuation is more likely to join. However, conditional on joining, the bidding incentive in the trading mechanism should depend only on the waiting cost but not the service value (cf. Hassin 1995). In general, mechanism design problems with multi-dimensional type distributions are hard even without queueing (cf. Rochet and Choné 1998), and one starter would be a model with perfectly correlated service values and waiting costs, as studied in Afèche and Pavlin (2016) for priority pricing. Even in this setting, the structure of the optimal mechanism is quite intricate.

Monotone virtual types. If virtual types are non-monotone (violating Assumption 1.2), then the service provider will not prefer strict priority, and the intermediary's optimal trading mechanism will have multiple pooling regions. Specifically, the optimal mechanisms would entail pooling customers in regions where the virtual types are locally decreasing, and pooling might also extend to the neighboring regions where the virtual types are increasing (Rayo 2013). Note that the social planner's preference for strict priority is immune to the properties of the virtual types.

M/M/1. If the service time does not have the memoryless property, then the preemptive trading rule will be problematic as the expected amount of waiting time traded would be unequal between the buyer in the queue and the seller at the server. This is a relatively easy

fix if we only allow trading to be non-preemptive, although the waiting time function will be slightly more complicated. If customers did not arrive according to a Poisson process, then the PASTA property would fail in determining both the expected waiting time transacted and the corresponding expected payment; the equilibrium behavior would be much harder to pin down. However, we expect most of the qualitative results to remain valid as the underlying economic forces of these results are not particularly tied to the $M/M/1$ assumption while the technical expressions are.

Some *behavioral aspects* of customers are noteworthy.

Endowment effect. The endowment effect (Kahneman et al. 1990) may be at work, engendering trading reluctance. If the waiting position is viewed as gains when acquired and as losses when sold, then loss aversion would imply potential sellers value their spot more so than potential buyers, ceteris paribus. The reluctance to sell one's spot would reduce the set of mutually agreed swaps in the market. El Hajia and Onderstal (2015) provide empirical evidence of endowment effects when human subjects trade queue positions.

Risk aversion. While our model assumes customers are risk-neutral toward both waiting times and monetary payments, real-world customers may have different risk preferences on these two dimensions. Specifically, Leclerc et al. (1995) show people tend to be more risk-averse about time than about money. The authors postulate that time is less fungible than money and therefore not as easily transferable or exchangeable. How these risk preferences affect customers' trading behaviors may be best tested in laboratory experiments.

Speculative behavior. One practical concern for introducing the trading marketplace is the arrival of "squatters", i.e., customers who game the system by selling their spots to garner money and renege when they have earned enough without actually receiving the service. These customers are typically time-insensitive and ascribe low valuation to the service itself and thus would not join the system otherwise. Another related phenomenon is "placeholders": those who are hired to wait in line for hours, only to be replaced at the very last moment by customers that truly request the service (cf. Gray 2009). On one

hand, the presence of speculators does not violate other customers' property rights since such swaps are still one-to-one substitution. On the other hand, these customers are likely to renege before entering the service, appropriating pecuniary gains that might otherwise be captured by the service provider. In principle, the up-front trade participation fee in the intermediary's optimal auction should deter some speculative customers. Additionally, the platform can act as a gatekeeper that closely monitors any suspicious trading activities and bans unscrupulous customers from using the trading platform if necessary. This further justifies the importance of the trading platform (intermediary) in mediating transactions. The understanding of speculative behaviors when customers trade queue positions is left for future research.

CHAPTER 2

SEARCH AMONG QUEUES UNDER QUALITY DIFFERENTIATION

2.1 Introduction

For customers seeking service, selecting a provider is rarely without hassle. Customers are typically not fully aware of the quality of a particular service provider; nor do they know for certain how long they have to wait before a provider is available. While quality of services naturally varies from one provider to another, the availability of services also differs due to short-run fluctuations of demand and supply. As a result, customers expend efforts garnering information about both quality and availability of services before making a choice. Such search frictions, costs incurred in finding a "good" service provider, are pervasive, for instance, in health care services.

Along with high patient demand, search frictions are believed to be contributing to long wait times for elective surgeries in many OECD (Organization for Economic Co-operation and Development) countries with public health systems, e.g. Canada, the United Kingdom (Sicilani and Hurst 2004, Siciliani and Hurst 2005). In the UK, 50,000 people waited longer than 18 weeks for elective surgical treatment in 2014 (Siddique 2015). In Canada, between referral from a general practitioner and receipt of treatment, survey results report a median waiting time of 18.3 weeks in 2015, 97% longer than in 1993 when it was 9.3 weeks (Barua 2015). Long wait times can have severe health, mental and economic consequences and thus are a major source of policy concern.

Among efforts to tackle excessive wait times, two policy initiatives have gained considerable traction. The first policy innovation is reducing search frictions of patients by providing more information to facilitate their choice over healthcare providers. For example, search is made easier nowadays in Canada by many provincial wait-time websites (WTA report card 2015) as well as a growing number of websites that provide quality ratings and reviews of

healthcare providers, such as RateMDs.com, the largest physician-rating site in Canada with more than 30 million unique visits per year as of 2014 (Kirkey 2014). In the UK, patients can look up National Health Services (NHS) Choices website for public hospital surgeons' safety records, waiting time, and patient reviews. Similarly, Healthcarereviews.com, a global doctor review website that include regions likes Canada and the UK, reviews doctors both on knowledgeability and wait times (Baum 2012). The second policy innovation is allowing privately funded health care to reduce the burden on public health systems (see e.g., Chen et al. 2015a). Countries like Germany, the Netherlands and Switzerland are known for primary/substitute private health insurance (PHI), whereas other countries, such as Australia, France and Canada, are home to complementary/supplementary PHI (Kreindler 2010).

Part of the rationale behind these two policies is that reducing search frictions or the system load can cut the overall wait times and consequently improve customer welfare for those who stay in the system. However, in a setting where service providers are vertically differentiated, it becomes less clear whether these policy interventions would succeed in reducing wait times. One could argue that customers might respond by crowding at high quality service providers instead of migrating toward ones with shorter waits. However, to what extent this argument is valid seems inconclusive. Even if a longer wait does arise, one would still expect customers to be better off on average, winding up with a "better" service provider after their search and customer welfare should strictly improve under both policy initiatives.

This chapter challenges this notion. We find that with either of these policy initiatives, not only may there be a longer wait, but the search reward (quality less waiting cost) customers reap from search may deteriorate, and individual customer welfare (search reward less total search costs) may *not* be improved. To this end, it is important to understand the following questions: (1) how do customers respond to these policy interventions: Do they search more or less? Are they prone to choosing service providers with a shorter wait or higher quality? (2) Given that waiting time is more measurable and often an explicit

policy target, what are the driving forces behind its change? (3) When would these perverse behaviors (a longer wait, worse search reward, and unimproved customer welfare) occur? We provide a framework to address these questions.

We consider a population of strategic, delay-sensitive customers who arrive to a queueing system with a collection of vertically differentiated parallel servers. Each customer conducts costly sequential search to resolve uncertainty about quality and queue length of servers. We note that while quality distribution is exogenously given, the queue length distribution is endogenously determined by customer search behavior. We characterize customers' search equilibrium in a tractable mean field model with an infinite number of servers. The equilibrium search strategies are of threshold type: for each of the servers' quality levels, a customer adopts a search threshold such that she joins a server of that quality level if the queue length is below the threshold, and continues to search otherwise. The higher the quality level, the higher the threshold. We show that there always exist either pure or mixed strategy equilibria of threshold type. In our model, we interpret the first policy proposal above (enhancing information provision) as a reduction in the search cost, and the second policy proposal (encouraging private practices) as a fall in the arrival rate.

We find that reducing the search cost induces a *load balancing effect* and a *quality substitution effect* on the average waiting time. With a lower search cost, customers adopt lower equilibrium search thresholds, sample more and seek a shorter queue for a given quality level. Hence, excessively long queues are eliminated, and this load balancing effect puts a downward pressure on the average waiting time. Meanwhile, however, customers substitute toward high quality servers for which they are willing to wait longer. This "quality substitution effect" puts an upward pressure on the average waiting time. In general, the overall reward of the two effects is ambiguous, and there are instances in which the second effect dominates the first one, thereby increasing the average waiting time. Moreover, the higher quality customers obtain may not offset the longer average waiting time, thus exacerbating customer *search reward* (the average customer utility upon joining) despite the fact that cus-

tomers conduct more search. Furthermore, reducing search costs may not improve customer welfare. Hence, expensive policy interventions may yield no returns to society.

When the arrival rate falls, lower system load puts an obvious downward pressure on the waiting time, but the load balancing effect and quality substitution effect are more nuanced. If the fall of the arrival rate is small enough to not change the equilibrium search thresholds, customers would sample less. When the original system load is low, we find that more customers end up joining low quality servers, so the quality substitution effect, like the lower system load, also exerts a downward pressure, and therefore the average waiting time is shorter on average. Nevertheless, when the system load is relatively high originally, the direction of the quality substitution effect is reversed, and its upward pressure on the average waiting time may be strong enough to overshadow the downward pressure from lower system load, thereby increasing the average waiting time. If the fall in the arrival rate changes the search thresholds, customers tend to sample more, yet in a manner that may further unbalance the system load, again pushing up the average waiting time and possibly decreasing the search reward. In addition, similar to the first policy intervention, the second one may yield no returns to individual customer welfare.

We provide conditions under which the two policy initiatives lengthen the average waiting time without improving customer welfare, and also conditions under which reducing search costs downgrade search rewards as well. We find that two key drivers for our results are substantive quality differentiation and relatively small search costs, both of which are characteristic of an environment that spawns strong search incentives. As the growing adoption of information technologies in recent years tends to create an environment of low search frictions, the two policy interventions may potentially backfire, presenting a new challenge for health authorities. It becomes increasingly important to take quality variations across providers into serious consideration before executing (costly) policy initiatives to further reduce search frictions or the system load.

This chapter makes three main contributions. First, to the best of our knowledge, our

framework is the first to take into account customers' three-way trade-off in quality, waiting and search costs. Second, we show that reducing the search cost or arrival rate makes a nuanced impact on the system, defying conventional wisdom. Third, we generate practically relevant insights and policy implications in the context of surgical waits in public health systems.

The remainder of this chapter is organized as follows. §2.2 surveys related literature. We set up the mean field model in §2.3. In §2.4, we establish the existence of equilibria. In §2.5 and §2.6, we conduct comparative statics analysis and examine the impact of the reducing search cost (corresponding to the policy initiative of disclosing more information) and reducing arrival rate (corresponding to the policy initiative of diverting demand to private health care), respectively. We conclude the chapter by discussing policy implications and future research directions in §2.7.

2.2 Related Literature

Our work straddles three streams of research: (i) search theory in economics and marketing, (ii) the supermarket model in queueing theory, (iii) queueing models with strategic customers.

2.2.1 The Economics of Search

The search theory literature is pioneered by Stigler (1961) in which the decision maker, faced with price dispersion (a probability distribution of prices), sets an a priori fixed sample size to minimize the expected total cost of purchase, including the search expenditure. The appropriateness of such fixed sample, simultaneous search models in describing rational consumer search behavior has been questioned (e.g., Nelson 1970) because consumers cannot commit to a fixed sample size in advance without adapting to new information as it comes in. Therefore, originating from McCall (1970) and Mortensen (1970), a large body of literature describes consumer behavior with *sequential search* models. A decision maker in a canonical

sequential search model takes a sequence of i.i.d. random draws from a reward distribution by incurring a (fixed) search cost per sample. After every sample, the decision maker decides whether to stop and earn the highest reward sampled so far, or, to incur additional search cost and sample another reward realization in the hope of observing a higher reward. The fundamental trade-off is thus: increasing the search reward at the expense of search costs. Our work builds on this sequential search literature both because of the economic rationality it captures and the analytical tractability it yields. The decision maker's welfare ("individual customer welfare") is the search reward upon stopping less the incurred search expenditure. The search expenditure is the search cost times the number of samples drawn before joining. From the literature (e.g., McCall 1970), we distill a number of intuitive principles of *optimal* search:

Principle 1: reducing search costs leads to sampling more servers before joining.

Principle 2: reducing search costs strictly increases individual customer welfare.

Principle 3: reducing search costs weakly increases the average search reward.

We note that in standard sequential search models, reducing search costs may either increase or decrease the search expenditure. The reward distribution determines the comparative static of the search expenditure (McCall and McCall 2008, pp. 80). As the comparative statics for individual customer welfare and the average search reward are unambiguous, we focus on these in this chapter.

Weitzman (1979) develops a sequential, directed search model in which the decision maker has some foreknowledge about the items under search, and thus also sets a selection rule in addition to the stopping rule. The sequential search framework of Weitzman (1979) has been widely employed in the marketing literature for structural estimation of customer search behavior on online platforms (Kim et al. 2010, Honka and Chintagunta 2017, Ursu 2016). The reader is referred to Rogerson et al. (2005) and McCall and McCall (2008) for surveys of search-theoretic models.

2.2.2 The Supermarket Model

The canonical model in this stream of literature studies a system in which each arriving customer draws a random sample of fixed size among a large collection of parallel servers (simultaneous search). The "reward" for the customer in this literature is joining a short queue, i.e. having low waiting cost. If the sample size is equal to one, there is essentially no search. When two or more queues are sampled, the strategy is to join the shortest queue in the sample. Increasing the sample size by one could be expensive. Conceptually, this amounts to introducing search costs. A mean field approach that passes the number of servers to infinity is commonly taken to analyze system performance. Both Vvedenskaya et al. (1996) and Mitzenmacher (2001) show in the mean field model that increasing the sample size from one to two significantly improves the expected reward upon joining (i.e. via a reduction of the waiting cost). The intuition is that when a customer joins the shortest queue among a larger sample, the system will be better "balanced". Hence the nomenclature "load balancing" for the search strategy described. Other seminal works include Turner (1998), Graham (2000) and Luczak and McDiarmid (2006). The tradeoff here is similar to that in the search literature: an increased reward is obtained at the expense of higher search cost. Therefore, the constructs in search theory (customer welfare, search reward and search expenditure) are also meaningful in the supermarket model. It is central to the supermarket model literature that the queue length distribution changes as a function of the customer search strategy (the number of samples drawn). Xu and Hajek (2013) endogenize the sample size in the supermarket model and formulate a mean field game among customers.

The present chapter differs from this literature in two important ways: (1) the "reward" in our model is jointly determined by both quality of servers and costs of waiting; (2) customers search sequentially rather than simultaneously. Basic queueing theory and the supermarket model literature indicate that one can either reduce the system load (a direct approach) or balance the system load with a larger sample size (an indirect approach) to increase customer welfare or the search reward. From the literature, we can obtain in a straightforward way

the following (intuitive) principles:

Principle 4: reducing the arrival rate (a) reduces the average waiting time, so, if all else equal, (b) strictly increases the average search reward and (c) strictly increases individual customer welfare.

Principle 5: controlling for the arrival rate, sampling more servers before joining (a) reduces the average waiting time (load balancing in Mitzenmacher 2001), and, hence, if all else equal, (b) strictly increases the average search reward and (c) strictly increases individual customer welfare.

2.2.3 Queueing Models with Strategic Customers

This literature stems from Naor (1969) who studies a single-server queue in which customers decide whether to join or balk based on the queue length observed. In Naor's model, all customers possess free queue length information, and queue choice is irrelevant since only one queue is available. Follow-up work extends Naor (1969) in various dimensions. Some consider choices among two parallel servers when queue length information is not freely revealed, e.g., Hassin and Haviv (1994), Hassin (1996). Others consider the joining decision in a single queue when customers have to pay an inspection cost to observe the queue (Hassin and Roet-Green 2016) or when there is information heterogeneity among customers (Hu et al. 2016).

Closely related to our work are Glazer and Hassin (1983), Davidson (1988) and Sattinger (2010), all of which study a setting where customers search among queues similar to the one studied in this chapter when servers have identical quality levels. Nevertheless, none of these papers establishes the existence of equilibria of the underlying queueing game. We prove the existence of either pure or mixed strategy equilibria in a more general framework where servers also differ in quality (and thus a vector of search thresholds needs to be determined in equilibrium).

Cui et al. (2016) study a game in which customers can retry a queue later (with a

retrial cost) if they decide neither to join nor to balk given the current queue length. They assume that the retrial interval is long enough such that the queue length distribution is uncorrelated across different retrials, whereas we focus on a system of an infinite number of servers such that queue length across servers is uncorrelated in the search process. Thus, their rational retrial model is similar in nature to our sequential search model when servers are homogeneous in quality. They show that when retrial cost is low such that there is no balking from customers, reducing retrial costs always strictly improves customer welfare, whereas we show that under quality differentiation, reducing search costs may not improve customer welfare.

Our work complements Hassin and Roet-Green (2015), who characterize the equilibrium of a sequential search model with two parallel servers with identical quality. They find that customers may follow a non-threshold strategy, joining the first server they inspect if its queue length is either short or long, but inspecting the second queue if the first server's queue length is intermediate. Customers would rather join the first server with a long queue than inspect the second server because of the correlation in queue length between the two servers. Customers infer from the long queue at the first server that the second queue may be even longer or not much shorter (otherwise previous customers would not join the first queue), thus not worth the cost of inspection. They refer to this behavior as "information cascades." We focus on a limiting system with an infinite number of servers, which induces threshold type equilibrium search strategies. The tractability of our mean field model enables us to study interesting comparative statics. We refer to Hassin and Haviv (2003) for detailed discussion on queueing models with strategic customers.

2.2.4 Contribution to the Literature

To the best of our knowledge, no previous work combines consumer search, queueing, and quality differentiation among servers. Hence, the five principles distilled from prior literature that would govern the impact of the two policy interventions under investigation (reducing

search costs and reducing the arrival rate) may not be adequate for such an important environment that arises naturally in practice. In this chapter, quality differentiation is the key driver to all our main results: without quality differentiation, we show that all five principles combined from the search and supermarket model literature carry over in a model with sequential equilibrium search. With quality differentiation, however, except for Principle 1, all principles may be violated. Therefore, it may be incomplete to assess policy interventions based on prior literature. The present chapter fills this gap and provides intuition for why combining insights from two the strands of literature is not sufficient to understand the impact of policy interventions in the presence of quality differentiation.

2.3 Model

We consider a system composed of n parallel servers (e.g., surgeons). The service time of each server follows an i.i.d. exponential distribution with mean $1/\mu$. Servers are vertically differentiated and each server's quality V is an i.i.d. draw from a discrete distribution with support $\{V_1, \ldots, V_N\}$ and $\mathbb{P}(V = V_i) = p_i \in (0, 1]$. $V_1 < V_2 < \cdots < V_N$. Let $\boldsymbol{V} = (V_1, \ldots, V_N)$ and $\boldsymbol{p} = (p_1, \ldots, p_N)$. First, Nature determines the quality of each server. Let the number of quality i servers be n_i. Then, delay-sensitive customers (e.g., patients) arrive to the system according to a Poisson process of rate $\Lambda = \lambda n$. Let $\rho = \frac{\lambda}{\mu}$ be the system load. For stability, we assume $\rho \in (0, 1)$. Services are provided on a first-come, first-served (FCFS) basis. Upon arrival, a customer does not know in advance the quality and queue length of any individual server, but is assumed to know the aggregate statistics about the system, $\boldsymbol{V}, \boldsymbol{p}, \lambda, \mu$.

Customers are expected utility maximizers and engage in costly sequential search to resolve uncertainty about quality and queue length, probing one server at a time uniformly at random without replacement. Each customer pays a search cost of $s > 0$ for sampling a server to observe both its quality i and queue length j. The search cost reflects the efforts customers have to exert in collecting information about both quality and queue length of a

server. This could be in terms of financial cost, e.g., making trips to hospitals to consult with general practitioners, or more pervasively, a psychic cost, i.e., the stress and anxiety involved in undertaking search and processing information. The sequential search framework allows customers to flexibly adapt their search/join decision to new information without committing to a fixed sample size in advance. We implicitly assume that one search query does not reveal either quality or queue length information of all servers. This is largely true even in online settings. For instance, despite the fact that multiple surgeons may be displayed with a single query on quality and wait time websites, the information presented at this level is usually very coarse. Patients still need to further evaluate each surgeon's quality by clicking on various reviews and ratings, and also double-check each surgeon's wait list given that it is dynamically evolving and thus may not be accurately reflected on these websites [1].

In the model, after each sample, customers decide whether to continue or terminate search. If a customer terminates search, she chooses to join, among all the servers she has sampled, the one with the highest expected reward (the best server). The expected reward of joining a server of quality i with queue length j is $V_i - c(j+1)/\mu$, where $c > 0$ is the waiting cost customers incur per unit time (including time spent at service). Since our focus is on customers' search/join behavior, we assume customers do not balk. In §3.4, we impose conditions on the model primitives to rationalize non-balking behavior.

While search, in general, is time consuming (and the search cost usually involves the opportunity cost of time incurred in search), here, for simplicity, we assume that search is instantaneous. This assumption may be somewhat reasonable in, e.g., slow-moving health service systems, where the time scale of searching for a surgeon (hours or days) may be somewhat negligible relative to the waiting time for elective surgeries that may amount to months or years. Instantaneous search eliminates the complication of changes in system

1. For example, the Ministry of Health in British Columbia claims on its provincial surgical website `https://swt.hlth.gov.bc.ca/` that, despite its best effort, "it cannot guarantee the completeness of the information as it is gathered from a variety of health authority sources."

states (queue length) when customers make join/search decisions[2]. It also ensures that customers join the system in the same order of their Poisson arrival. Moreover, instantaneous search immediately implies it is sufficient for customers in the search process to consider the best server ever sampled as a potential candidate to join instead of keeping track of all samples. In the model, customers' expected utility if they decide to search depends on their rational expectation of queue length distribution in the system, which, in turn, depends on other customers' search strategies. This constitutes a queueing game among self-interested customers.

When the number of servers in the system is finite, finding Nash equilibrium of this game is a very cumbersome problem even when there are only two servers with identical quality, as shown in Hassin and Roet-Green (2015) (refer to the discussion in §2.2.3). The information cascades in a finite system would result in non-threshold search strategies. Nevertheless, when the number of servers grows large, the problem becomes *more* tractable analytically. As one takes a sample from a large collection of servers, knowing the queue length of one server provides little information about the queue length of her next search. Therefore, customers' search history will not alter what they anticipate if they continue to search in a steady state system. Thus, the aforementioned information cascades are no longer present, and customers revert to threshold strategies, by which they join a queue if its queue length is below a certain threshold, and search otherwise. Hence, we are able to characterize analytically the equilibrium search and establish some counterintuitive comparative static results that would not be otherwise possible.

For our equilibrium characterization and analysis, we focus on a mean field model in which the number of servers tends to infinity, i.e., $n \to \infty$. Applying the mean field approach (e.g., Mitzenmacher 2001), we can characterize the mean field equilibrium that serves as a tractable large market approximation to a finite system. In 2.3.1, we first conjecture that

2. Otherwise, customers may find the queues they sampled evolve as they search, thus having an incentive to sit on a server already probed, i.e., neither join nor search, hoping to join it later when its queue length gets shorter.

customers follow a symmetric threshold strategy (which we specify shortly) in the limiting process with $n \to \infty$. For a given threshold strategy, we derive the stationary queue length distribution. We later verify in §2.3.2 that given any stationary queue length distribution, customers' best response is indeed a threshold strategy in the limiting process.

2.3.1 Stationary Queue Length Distribution

In this subsection, we derive the stationary queue length distribution in the mean field model given any customer search strategy of threshold type. Let $\boldsymbol{k} = (k_1, ..., k_N) \in \mathbb{N}^N$ denote the "search thresholds" in customers' strategy whereby a server of quality i is accepted if and only if its queue length (including the customer in service) is strictly less than k_i; otherwise, customers continue to search. In particular, $k_i = 0$ means that customers never join a server of quality i, even if that server is idle upon arrival. Since no customers join a server of quality i if there are already k_i customers waiting in line, the maximum queue length observable is simply k_i. This implies that customers continue to search if and only if the current queue length is k_i for server of quality i. It should be emphasized that the system evolution is a function of customers' search strategy, $\boldsymbol{k}$. Denote the number of servers of quality i at time t with a queue length of j by $n_j^i(t, \boldsymbol{k})$. Define $\pi_j^i(t, \boldsymbol{k}) = n_j^i(t, \boldsymbol{k})/n_i$ to be the fraction of servers of quality i at time t with a queue length of j. The state of the system at any time t can be represented by an N dimensional vector $\boldsymbol{\pi}(t, \boldsymbol{k}) = (\boldsymbol{\pi}^1(t, \boldsymbol{k}), ..., \boldsymbol{\pi}^N(t, \boldsymbol{k}))$, where each component is itself a $k_i + 1$ dimensional vector, i.e., $\boldsymbol{\pi}^i(t, \boldsymbol{k}) = (\pi_0^i(t, \boldsymbol{k}), \pi_1^i(t, \boldsymbol{k}), \ldots, \pi_{k_i}^i(t, \boldsymbol{k}))$. For given[3] $(n_1, \ldots, n_N)$ and $\boldsymbol{k}$, the system evolution constitutes a Markov Chain on the state space above. We refer to $\boldsymbol{\pi}(t, \boldsymbol{k})$ as the queue length distribution at time t given strategy $\boldsymbol{k}$.

We formulate a set of ordinary differential equations (ODEs) to characterize the time evolution of the limiting process with $n \to \infty$, and we solve for the fixed point of the

3. n_i is a fixed number (not a random variable) since Nature determines the quality of each server before customers arrive.

limiting process, which gives the stationary queue length distribution of the system, $\pi_j^i(\boldsymbol{k})$, $0 \leq j \leq k_i,\ 1 \leq i \leq N$. Note that $\boldsymbol{\pi}(\boldsymbol{k})$ is a random vector in a finite system even in steady state; at the fixed point of the limiting system, however, it is deterministic (Mitzenmacher (2001)). Lemma 2.1 provides the expression for the fixed point of $\boldsymbol{\pi}(\boldsymbol{k})$. and the detailed steps of derivation are found in the proof relegated to Appendix B.2.

Lemma 2.1. *For $\boldsymbol{k} = (k_1, \ldots, k_N) \geqslant \mathbf{0}$, define the index set $\mathcal{S} = \{i | k_i > 0\}$. If $\sum_{i \in \mathcal{S}} p_i > \rho$, the stationary queue length distribution of the limiting process is given by:*

$$\pi_j^i(\boldsymbol{k}) = \frac{[\alpha(\boldsymbol{k})]^j}{\sum_{l=0}^{k_i} [\alpha(\boldsymbol{k})]^l} \quad \text{for } 0 \leqslant j \leqslant k_i,\ i \in \{1, \ldots, N\}, \tag{2.1}$$

where $\alpha(\boldsymbol{k})$ is the (unique) positive root α of the following equation:

$$\frac{\alpha}{\rho} = \frac{1}{1 - \sum_{i=1}^{N} p_i \frac{\alpha^{k_i}}{\sum_{l=0}^{k_i} \alpha^l}}. \tag{2.2}$$

The expression of $\pi_j^i(\boldsymbol{k})$ in (2.1) suggests that each server of quality i acts like an independent $M/M/1/k_i$ system with arrival rate $\alpha(\boldsymbol{k})\mu$ and service rate μ, where $\alpha(\boldsymbol{k})$ is endogenously determined by (2.2). As such, one can interpret $\alpha(\boldsymbol{k})\mu$ as the rate at which customers inspect any given server. This rate is the same for all servers due to undirected, random search. We refer to $\alpha(\boldsymbol{k})$ as "search intensity." The right hand side of (2.2) is the average number of samples a customer draws before joining a particular server. This follows from the fact that the number of samples is a geometrically distributed random variable with the probability of observing a "full" queue (that customers do not join) being $\sum_{i=1}^{N} p_i \frac{\alpha^{k_i}}{\sum_{l=0}^{k_i} \alpha^l}$. According to Equation (2.2), search intensity $\alpha(\boldsymbol{k})$ is such that the rate at which customers inspect any given server, $\alpha(\boldsymbol{k})\mu$, divided by the rate at which customers arrive, $\rho\mu$ should be equal to the average number of servers each customer samples. The condition $\sum_{i \in \mathcal{S}} p_i > \rho$ on the search strategy $\boldsymbol{k}$ ensures that the underlying system is stable.

2.3.2 Best Response

We now derive customers' best response given any arbitrary stationary queue length distribution $\boldsymbol{\pi}$ in the mean field model (not necessarily generated by the search strategy $\boldsymbol{k}$ in §2.3.1). Specifically, by stationary queue length distribution, we require that π_j^i, the fraction of servers of quality i with queue length j, be a fixed number that does not change over time for all i and j. This would not be possible in a system with finite servers because π_j^i is a random variable even in steady state. We focus on the case in which all customers, who are ex-ante symmetric, follow identical strategies.

As we argued earlier, each customer only needs to keep track of the best server ever sampled, and takes an action $a \in \{\text{join}, \text{search}\}$ each time after she samples a new server. If she chooses $a = \text{join}$, she joins that best server. If she chooses $a = \text{search}$, she pays a search cost s, and takes another sample. Each customer maximizes her expected utility, $U(i, j; \boldsymbol{\pi})$, given quality i and queue length j of the best server and distribution $\boldsymbol{\pi}$. Thus, a pure strategy is a mapping $\sigma : \{1, \ldots, N\} \times \mathbb{N} \mapsto \{\text{join}, \text{search}\}$, where $\mathbb{N}$ is the set of nonnegative integers representing any possible queue length. Each customer's decision problem can be formulated as a dynamic program with the following Bellman equation:

$$U(i, j; \boldsymbol{\pi}) = \max \begin{cases} V_i - c(j+1)/\mu, & a = \text{join} \\ -s + \sum_{i'=1}^{N} p_{i'} \sum_{j'=0}^{\infty} \pi_{j'}^{i'} U(i', j'; \boldsymbol{\pi}), & a = \text{search} \end{cases}. \tag{2.3}$$

This is a classical optimal stopping problem, and from search theory (e.g. McCall and McCall 2008), the optimal strategy is of threshold type defined as follows. Let $\bar{U}$ denote each customer's expected utility from search (the second line in (2.3)). Thus, a customer joins a server of quality i and queue length j if the expected reward of joining that server, $V_i - c(j+1)/\mu$, is at least as large as $\bar{U}$; otherwise, she continues to search. We refer to this threshold $\bar{U}$ as the "reservation utility," the amount utility customers expect to derive from the system. Because of the threshold structure, customers do not even need to hold

on to the best server ever found and recall it later when search is terminated; it suffices to only evaluate the most recently sampled server (the current server) against $\bar{U}$. Since $\bar{U}$ is fixed, a server not acceptable now will not be acceptable later. We reiterate that this threshold structure to hold crucially hinges on the queue length distribution $\boldsymbol{\pi}$ being fixed. This condition would not generally be satisfied in a finite-server system because correlation across queues would enable customers to infer the queue length of the next sample from their search history by updating their belief about $\boldsymbol{\pi}$ (which is stochastic).

We can equivalently represent the optimal threshold strategy by specifying a search threshold k_i for every quality i such that each customer rejects the current server and keeps on searching if and only if its queue length $j \geq k_i$. Thus, by definition, $\bar{U} \in [V_i - c(k_i+1)/\mu, V_i - ck_i/\mu]$. Specifically, if $k_i = 0$, i.e., customers continue to search even when a server of quality i is idle, $\bar{U}$ only needs to satisfy $\bar{U} \geq V_i - c/\mu$. Let $\boldsymbol{k} = (k_1, \ldots, k_N) \in \mathbb{N}^N$ be such thresholds, then the reservation utility of a customer who follows strategy $\boldsymbol{k}$ given distribution $\boldsymbol{\pi}$ is

$$\bar{U}(\boldsymbol{k};\boldsymbol{\pi}) = -s + \sum_{i=1}^{N} p_i \left[\sum_{j=0}^{k_i-1} \pi_j^i (V_i - c(j+1)/\mu) + \sum_{j=k_i}^{\infty} \pi_j^i \bar{U}(\boldsymbol{k};\boldsymbol{\pi}) \right],$$

or, after rewriting,

$$\underbrace{\bar{U}(\boldsymbol{k};\boldsymbol{\pi})}_{\triangleq CW(\boldsymbol{k},\boldsymbol{\pi})} = -s \underbrace{\frac{1}{1 - \sum_{i=1}^{N} p_i \sum_{j=k_i}^{\infty} \pi_j^i}}_{\triangleq ES(\boldsymbol{k};\boldsymbol{\pi})} + \underbrace{\frac{\sum_{i=1}^{N} p_i \sum_{j=0}^{k_i-1} \pi_j^i (V_i - c(j+1)/\mu)}{1 - \sum_{i=1}^{N} p_i \sum_{j=k_i}^{\infty} \pi_j^i}}_{\triangleq ER(\boldsymbol{k};\boldsymbol{\pi})}. \tag{2.4}$$

Since the reservation utility $\bar{U}(\boldsymbol{k};\boldsymbol{\pi})$ is the utility customers expect to derive from the system, we refer it as *individual customer welfare*, denoted by $CW(\boldsymbol{k};\boldsymbol{\pi})$. It is equal to the average *search reward*, $ER(\boldsymbol{k};\boldsymbol{\pi})$, less the *search expenditure*, $sES(\boldsymbol{k};\boldsymbol{\pi})$. $ES(\boldsymbol{k};\boldsymbol{\pi})$ is the average number of samples each customer draws. The average search reward, $ER(\boldsymbol{k};\boldsymbol{\pi})$, measures how "good" of a service provider customers get on average.

As in Naor's model, in which the joining "reward" is the "quality" of the server less the waiting cost, the average search reward in our model, $ER(\boldsymbol{k};\boldsymbol{\pi})$, is the average quality customers obtain, $EV(\boldsymbol{k};\boldsymbol{\pi})$ (as in the search literature) less the average waiting cost of the selected server $cEW(\boldsymbol{k};\boldsymbol{\pi})$ (as in the supermarket model), where $EW(\boldsymbol{k};\boldsymbol{\pi})$ is the average waiting time:

$$ER(\boldsymbol{k};\boldsymbol{\pi}) = \underbrace{\frac{\sum_{i=1}^{N} p_i V_i \sum_{j=0}^{k_i-1} \pi_j^i}{1-\sum_{i=1}^{N} p_i \sum_{j=k_i}^{\infty} \pi_j^i}}_{EV(\boldsymbol{k};\boldsymbol{\pi})} - c \underbrace{\frac{\sum_{i=1}^{N} p_i \sum_{j=0}^{k_i-1} \pi_j^i (j+1)/\mu}{1-\sum_{i=1}^{N} p_i \sum_{j=k_i}^{\infty} \pi_j^i}}_{EW(\boldsymbol{k};\boldsymbol{\pi})}.$$

It is possible for a system with higher individual customer welfare to deliver lower average search reward, if the improvement in welfare is driven by savings in search expenditure. Notice that individual customer welfare is equal to the reservation utility. Thus, individual customer welfare is a measure of the *worst* expected utility *upon joining*, whereas the average search reward is a measure of the *average* expected utility *upon joining*. Both measures may be of interest to policymakers.

For a given $\boldsymbol{\pi}$, we define the Best Response, $BR(\boldsymbol{\pi})$, to be the (set of) optimal search thresholds:

$$\begin{aligned} BR(\boldsymbol{\pi}) \triangleq \{\boldsymbol{k} \in \mathbb{N}^N : \bar{U}(\boldsymbol{k};\boldsymbol{\pi}) \in [V_i - c(k_i+1)/\mu, V_i - ck_i/\mu] \text{ if } k_i > 0; \\ \bar{U}(\boldsymbol{k};\boldsymbol{\pi}) \geq V_i - c/\mu \text{ if } k_i = 0, \quad \forall i = 1,\ldots,N\}, \end{aligned} \tag{2.5}$$

where $\bar{U}(\boldsymbol{k};\boldsymbol{\pi})$ is defined in (2.4).

2.4 Equilibrium

In the mean field model, on one hand, any conjectured (stable) threshold search strategy produces a stationary queue length distribution (fixed point) by §2.3.1; on the other hand, by §2.3.2, the best response to any stationary queue length distribution is indeed a threshold strategy. Therefore, it is natural to look for mean field equilibria of the queuing game that

involve threshold strategies. A mean field Nash equilibrium consists of strategies represented by search thresholds $\boldsymbol{k}^* = (k_1^*, k_2^*, ..., k_N^*)$ that are the best responses to the stationary queue length distribution they induce, i.e., the equilibrium satisfies a consistency check. Formally, $\boldsymbol{k}^* \in \mathbb{N}^N$ is a symmetric mean field pure strategy Nash equilibrium if

$$\boldsymbol{k}^* \in BR(\boldsymbol{\pi}(\boldsymbol{k}^*)), \tag{2.6}$$

where $BR(\cdot)$ is defined in (2.5). Recall that we focus on *symmetric* equilibria since customers are ex-ante identical; we study a limiting system where both the number of servers and the arrival rate are scaled to infinity, and hence the notion *mean field*.

Conditions (2.6) only specify pure strategy equilibria. We defer the definition of mixed-strategy equilibria to §2.4.1. To ensure that it is not rational for customers to balk (for which they get zero utility), we make the following assumption on the model primitives throughout the chapter.

Assumption 2.1. *$V_i - c/\mu > 0$ for $i = 1, \ldots, N$. $\sum_{i=1}^N p_i \sum_{j=0}^{\bar{k}_i - 1} \pi_j^i(\bar{\boldsymbol{k}})(V_i - c(j+1)/\mu) > s$ with $\bar{\boldsymbol{k}} = (\bar{k}_1, ..., \bar{k}_N)$ and $\bar{k}_i = \lfloor V_i \mu / c \rfloor$ for $i = 1, \ldots, N$.*

Note that $\bar{k}_i$ is the Naor threshold for quality i (Naor 1969). Assumption 2.1 says that if everyone adopts the Naor search thresholds, each customer's expected utility is still positive. This guarantees the equilibrium search thresholds be bounded above by the Naor thresholds. Let $d_i \triangleq (V_i - V_1)\,\mu/c$. Let $\lfloor x \rfloor$ be the largest integer less than or equal to x; $\lceil x \rceil$ be the smallest integer greater than or equal to x. Lemma 2.2 immediately follows from Conditions (2.5).

Lemma 2.2. *For any pure-strategy equilibrium $\boldsymbol{k}^*$, there exists an $M \in \{1, \ldots, N\}$ such that $k_{i+1}^* - k_i^* \in \{\lfloor d_{i+1} - d_i \rfloor, \lceil d_{i+1} - d_i \rceil\}$, $\forall i = M, \ldots, N-1$ and $k_i = 0$ for $i < M$. Particularly, if $M = 1$, then $k_i \geq 1$ for all $i = 1, \ldots, N$.*

Lemma 2.2 suggests that the difference between the equilibrium search thresholds of two quality levels is roughly proportional to the difference between the quality levels themselves.

Servers below quality level M are screened out. Note that the higher the search threshold, the less selective customers are. Thus, the equilibrium structure is intuitive: with a higher quality comes a higher search threshold, and customers are willing to accept a longer queue. Notice that the absolute values of the quality do not have any material consequence on the equilibrium search thresholds, but the relative values (d_i) do. The rationale is that customers either join or search (yet do not balk), and, therefore, if the expected utilities from both joining and searching are equally changed, the two effects will offset each other, leaving customers' decision unaltered. If the parameters $(\mu, c, \boldsymbol{V})$ are such that $d_i \in \mathbb{N}, \forall i$, then $k^*_{i+1} - k^*_i = d_i - d_{i+1}, \forall i = M, \ldots, N-1$, which implies for each conjectured k_M, $k_i = k_M + d_i - d_M, \forall i = M, \ldots, N-1$. Thus, by substitution, solving for the equilibrium for a given M reduces to a one-dimensional fixed point problem (over k_M, for example). Of course, M also has to be determined. In general, d_i may be non-integer, which entails a multidimensional search for $\boldsymbol{k}$. Despite this technical challenge, we establish the existence of equilibria in §2.4.1.

2.4.1 Existence of Equilibria

As pure strategy equilibria may not always exist in general, we expand our strategy space to incorporate mixed (threshold) strategies that customers may play. Following the notational convention in Hassin and Haviv (1997), we denote a mixed threshold strategy for quality i by $k_i = j + \kappa_i$, $\kappa_i \in (0, 1]$. A customer who plays strategy k_i always joins the queue if the queue length is at most $j-1$, always rejects the queue and keeps searching if the queue length is at least $j+1$, and randomizes between the two with a joining probability κ_i when the queue length is j. If $\kappa_i = 1$, then k_i reduces to a pure strategy search threshold. In Appendix B.1, we provide a constructive proof of existence of either pure or mixed strategy equilibria. Our construction allows us to determine all pure and mixed strategy equilibria in the game. The key idea is translating the multi-dimensional fixed-point problem over search thresholds $\boldsymbol{k}$ to a single dimensional fixed-point problem over the reservation utility $\bar{U}(\boldsymbol{k}; \boldsymbol{\pi}(\boldsymbol{k}))$.

Theorem 2.1 (Existence and Refinement of Equilibria)**.** *There exists either a pure or mixed strategy threshold type equilibrium: of all the equilibria, the one with the highest expected utility* $\bar{U}(\boldsymbol{k}^*; \boldsymbol{\pi}(\boldsymbol{k}^*))$ *is the Pareto-dominant equilibrium.*

Since multiple equilibria may arise, we select the one that yields the highest $\bar{U}(\boldsymbol{k}^*; \boldsymbol{\pi}(\boldsymbol{k}^*))$ as a means to refine the equilibrium concept. While each customer's ex-post utility depends on the particular server they sample, $\bar{U}(\boldsymbol{k}^*; \boldsymbol{\pi}(\boldsymbol{k}^*))$ is the utility each customer expects from the system before they sample any server. Thus, the equilibrium with the highest $\bar{U}(\boldsymbol{k}^*; \boldsymbol{\pi}(\boldsymbol{k}^*))$ is a Pareto-dominant equilibrium because customers (who are identical) would be better off ex-ante in this equilibrium. This refinement criterion also serves to pin down a *unique* equilibrium. We focus on the Pareto dominant equilibrium in our subsequent analysis.

Pure strategy equilibria may not exist under multiple quality levels ($N \geq 2$) and in those cases we have to resort to mixed strategy equilibria. However, when there is no quality differentiation, Theorem 2.2 below shows that pure strategy equilibria always exist and moreover, that one of them is the Pareto dominant equilibrium.

Theorem 2.2. *For* $N = 1$*, There always exists a Pareto-dominant pure-strategy equilibrium.*

When there is no quality differentiation, mixed strategy equilibria will not emerge once the refinement criterion is applied, and relying on this result, we can show all five principles discussed above hold (see Appendix B.3). We shall see in §2.5 and §2.6 that some of our key results rely on the presence of mixed strategy equilibria when servers are vertically differentiated.

From now on, we denote the equilibrium system metrics by $X^* = X(\boldsymbol{k}^*; \boldsymbol{\pi}(\boldsymbol{k}^*))$ for $X = CW, ES, ER, EV$ and EW.

Proposition 2.1. *In equilibrium,* $EV^* \geq \sum_{i=1}^{N} p_i V_i$.

Proposition 2.1 confirms the intuition that due to search, customers expect to obtain a higher quality of service than the average quality of servers in the system (without search).

Since customers adopt a larger search threshold with high quality servers, this more lenient search criterion implies customers are less likely to screen out a high quality server sampled. Therefore, more customers end up being served by high quality servers, and the average quality obtained is improved.

2.5 Impact of Policy Intervention 1: Search Cost Reduction

In this section, we analyze how reducing search costs affects customer and system behavior in equilibrium. We report the impact on individual customer welfare, the average sample size, the average quality obtained, the average waiting time, the average search reward and search expenditure.

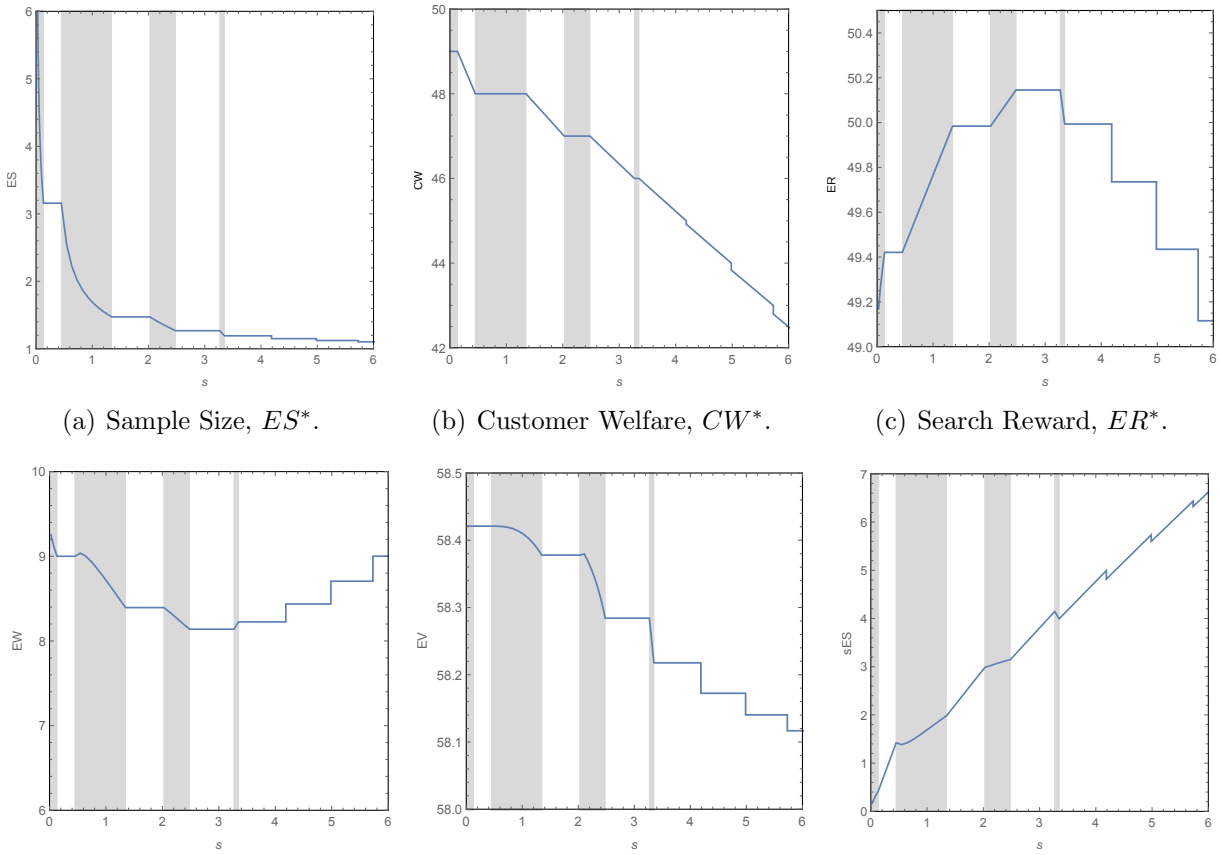
2.5.1 An Illustrative Example

Example 2.1. *Consider the following parameters:* $N = 2$, $\mu = 1$, $\lambda = 0.95$, $c = 1$, $\boldsymbol{p} = (0.2, 0.8)$, $\boldsymbol{V} = (50, 60)$. *The system metrics versus the search cost are plotted in Figure 2.1.*

Figure 2.1-(a) illustrates that lowering search costs leads to more sampling, which confirms Principle 1. Figure 2.1-(b) illustrates that individual customer welfare might fail to strictly increase as search costs fall, and this is a violation of Principle 2. Figure 2.1-(c) shows that the average search reward may deteriorate with a reduction in search costs, which contradicts Principle 3. Figure 2.1-(d) indicates that the average waiting time may increase with more sampling (observed from Figure 2.1-(a)), which conflicts with Principle 5. Figure 2.1-(e) illustrates that customers obtain higher quality as search costs fall. Figure 2.1-(f) suggests that due to lower search costs and more sampling, search expenditure is, in general, non-monotone in search costs.

To understand what is driving the invalidity of Principles 2, 3, 5, we dissect the change of search costs into three cases. First, if changing the search cost does not result in the

Figure 2.1: System metrics versus search cost for Example 2.1.



(a) Sample Size, ES^*. (b) Customer Welfare, CW^*. (c) Search Reward, ER^*.

(d) Waiting Time, EW^*. (e) Quality, EV^*. (f) Search Expenditure, sES^*.

Note. Mixed-strategy equilibria are indicated by shaded areas.

change of the (pure strategy) equilibrium search thresholds, then EV^*, EW^*, ES^*, ER^* are unaffected. Both individual customer welfare and the search expenditure are linear in search cost with a slope equal to the average sample size ES^* under that equilibrium. The improvement in individual customer welfare stems solely from the reduction of search cost per probe.

Second, one pure-strategy equilibrium moves to another with a lower search cost, there is a discontinuous upward jump in individual customer welfare and the average sample size. A reduction in search costs leads to an increase in the search expenditure. Both the average search reward and waiting time improve. These jumps occur at relatively large search cost

levels ($s > 4$ in Figure 2.1).

Third, if a fall in search costs shifts the system from one mixed-strategy equilibria to another, then individual customer welfare CW^* remains unchanged. Customers sample more, but it becomes ambiguous in which direction the search expenditure sES^* moves. In Figure 2.1-(f), the search expenditure can be either decreasing or increasing in those shaded areas indicating mixed strategy equilibria. In case of *inelastic* search, i.e., the rate of increase in the amount of sampling does not catch up with the rate of decrease in search costs such that the search expenditure declines. The average search reward ER^* deteriorates (since $ER^* = sES^* + CW^*$) and wait times tends to get longer. The improved quality customers get in those instances does not make up for the prolonged wait, diminishing search reward despite more search. These phenomena (violating Principles 2 3, 5) occur under relatively small search costs ($s < 2.5$ in Figure 2.1).

In summary, reducing search costs may not improve customer welfare, yet strictly lengthen the average waiting time and degrade the average search reward. These findings are in disagreement with Principles 2, 3, 5. We explain these numerical findings with our analytical development.

2.5.2 Analytic Results

Equipped with the numerical observations above, we now identify sufficient conditions on the model primitives under which decreasing search costs yields no improvement in customer welfare, increases the average waiting time and degrades the search reward.

Proposition 2.2. *Consider $N \geq 2$. If $(1 - p_1)\mu < \lambda$, there exists an $\bar{s}$ such that for any $s < \bar{s}$, customers play a mixed strategy equilibrium with search threshold $k_1^* \in (0, 1)$ for the lowest quality level and $k_i^* = 1 + \lfloor d_i \rfloor$ for $i = 2, \ldots, N$. As search cost s decreases within $(0, \bar{s})$,*

(i) customers sample more servers before joining, i.e., ES^* increases;

(ii) individual customer welfare CW^* remains unchanged;

(iii) customers obtain higher quality on average, i.e., EV^* increases;
(iv) the average waiting time EW^* increases if $\lfloor d_N \rfloor \geq 1$ and remains unchanged otherwise;
(v) when $d_i \in \mathbb{N}_+ \ \forall i$ and arrival rate λ is high enough, the average search reward ER^* decreases.

Under the conditions specified in Proposition 2.2, customers play a mixed-strategy, joining an empty server of the lowest quality level with a certain probability. They do not randomize at other quality levels. Customers are highly selective in their search (sometimes not even joining an empty server of low quality); they can afford to do so due to low enough search costs: $s < \bar{s}$. As search costs further decrease, customers sample even more (Proposition 2.2-(i)), which confirms Principle 1.

By virtue of the mixed-strategy equilibrium, customers are indifferent between search which gives the reservation utility (average customer welfare), and joining an empty low quality server which gives an expected utility $V_1 - c/\mu$. Hence, individual customer welfare is unchanged with search costs when $s \in (0, \bar{s})$ and equal to $V_1 - c/\mu$ (Proposition 2.2-(ii)), which is in direct violation of Principle 2. In other words, investing in a reduction in search costs would yield *no* improvement in customer welfare. Notice that two unique features of our model drive this result: quality differentiation and equilibrium search. Without quality differentiation, there would be no need for randomization in a Pareto dominant equilibrium (see Theorem 2.2). Without equilibrium search, customers could be strictly better off with a lower search cost by sticking to the original (randomization) search strategy; yet, everyone has an incentive to deviate from their original strategy, resulting in a different reward distribution (which is a combination of the quality and queue length distribution) that brings equilibrium customer welfare back to the original level. As a consequence of such competitive interplay between customers, only a mixed strategy can be supported in equilibrium, leading to the result that customer welfare does not improve when search costs fall.

While Proposition 2.2-(i) suggests customers generally conduct more sampling, Proposi-

tion 2.2-(iii) goes on to show that customers are more selective in a particular way that leads them to join an empty, server of the lowest quality with a smaller probability. Therefore, more customers end up joining higher quality servers, improving the overall quality obtained. We refer to this phenomenon as the *quality substitution effect*.

As Proposition 2.2-(iv) suggests, a consequence of the quality substitution effect is that the average waiting time strictly increases as long as quality differentiation is not too trivial ($\lfloor d_N \rfloor \geq 1$)[4]. This is because customers are willing to form a longer queue in return for higher quality. Recall from Proposition 2.2-(i) that customers sample more when search costs fall (confirming Principle 1). Proposition 2.2-(iv) thus contradicts Principle 5(a): despite more sampling, customers expect increased waiting time. Here, more sampling un-balances the system load since customers' incentive to choose a shorter queue is subdued by the alternative desire for higher quality.

At this point, it is still undetermined how the search reward changes because higher quality is experienced in conjunction with a longer wait. Proposition 2.2-(v) shows that if the arrival rate is significant (a situation particularly relevant for real world applications of our interest), the increase in the average waiting time outweighs the improvement in quality obtained, diminishing search reward despite more search. We show this by showing search becomes inelastic, i.e., sES^* decreases as s falls. Since $ER^* = CW^* + sES^*$ and CW^* is not changed, less search expenditure here implies worse search reward. Proposition 2.2-(v) is in direct violation of Principle 3.

The key condition in Proposition 2.2 is search cost s being small enough. As improved technologies keep alleviating search frictions, it is likely that real-world system eventually encounters the range $(0, \bar{s})$, which implies that a longer waiting time as a result of lower search costs may be especially relevant and hence a real concern. Moreover, we show in the appendix that $\bar{s}$ is increasing in integer d_N, which implies that a system with greater quality

4. We require $\lfloor d_N \rfloor \geq 1$ because if $\lfloor d_N \rfloor = 0$, customers only join empty servers at all quality levels and the average waiting time is constantly $1/\mu$, unaffected by search cost $s \in (0, \bar{s})$.

differentiation is more likely to exhibit this behavior even under relatively large search cost levels. Finally, it is required that if the lowest quality servers is left out, there is not enough capacity to serve the entire market ($(1-p_1)\mu < \lambda$). This condition speaks to the need of quality heterogeneity in driving the result. If $N = 2$, for example, $1 - p_1 = p_2$; higher values of p_2 and low search costs would lead to all customers exclusively joining high-quality servers, effectively reducing the problem to the homogeneous quality case in which all principles hold.

Note that the conditions identified in Proposition 2.2 are only sufficient and not necessary. They represent a small slice of the parameter space (e.g., the first shaded area from the left in Figure 1) where we can sharply characterize the directional changes of the system metrics. As illustrated in Figure 1, our results hold in a wide range of regions when conditions in Proposition 2.2 do not apply ($s > \bar{s}$). Nevertheless, in those other regions, identifying sufficient conditions for our results is more involved and those conditions may also be too complex to be useful. We do not attempt this, but, rather, reveal in the sequel the driving forces that underlie the change in search costs. Proposition 2.3 studies the comparative statics of a series of system metrics under pure strategy equilibria.

Proposition 2.3. *If both $\boldsymbol{k}^*$ and $\boldsymbol{k}^{*\prime}$ are pure strategy equilibria under s and s', respectively, where $s < s'$, then,*

(i) customers adopt lower thresholds under s than under s', i.e., $\boldsymbol{k}^* \leq \boldsymbol{k}^{*\prime}$ component-wise;

(ii) customers sample more under s than under s', i.e., $ES^* \geq ES^{*\prime}$;

if $k_i^{*\prime} - k_i^*$ is constant in i for $i = 1, \ldots, N$,

(iii) customers obtain higher quality on average under s than under s', i.e., $EV^* \geq EV^{*\prime}$.

Proposition 2.3-(i) and (ii) confirms Principle 1 in a more general context that a lower search cost would prompt customers to sample more by accepting shorter queues only, i.e., adopting lower search thresholds. Recall that the search threshold for a given quality level is also the maximum possible queue length of servers with that quality level. Hence, smaller search thresholds eliminate long queues and impose a downward pressure on the average waiting time. This is reminiscent of Principle 5 and we call this the *load balancing effect*.

Nevertheless, Proposition 2.3-(iii) suggests that there is also a quality substitution effect at work (as in Proposition 2.2) whereby customers end up joining higher quality servers on average. As a general force that governs customer search behavior, the quality substitution effect can be viewed as an extension of Proposition 2.1, which only says that customers gravitate toward high quality servers if given the option to search. By comparison, Proposition 2.3-(iii) takes one step further by indicating that such quality substitution persists as the option to search becomes cheaper. The rationale is the following. Lower search costs eliminate long queues, so queue length is redistributed among servers. Since low quality servers have shorter queues, i.e., fewer queue length possibilities, such redistribution will increase the fraction of "full" queues (that customers would reject) more so at low quality servers than at their high quality counterparts. As a consequence, this increases the likelihood of customers eventually joining higher quality servers. We show in the proof of Proposition 2.3 that the distribution of quality associated with lower search costs is stochastically larger than the distribution of quality associated with higher search costs, and therefore, customers get access to better quality of service on average. This is an interesting result in service systems because it is not a priori clear whether customers migrate toward shorter queues or toward high quality service providers with a lower search cost.

The quality substitution effect implies that a larger fraction of the customer population tends to be served by higher quality servers, for which they may wait longer (since customers adopt a higher search threshold with servers of higher quality). Consequently, the quality substitution effect puts an upward pressure on the average waiting time, counteracting the downward pressure from the load balancing effect. The direction of the overall effect is ambiguous in general, as shown in Figure 2.1-(d). However, despite the complicated dynamics, recall from Proposition 2.2 that when the search cost is sufficiently small, only the quality substitution effect is present, and the average waiting time will unequivocally increase when search costs fall.

Proposition 2.4. *In general, as search cost s decreases,*

(i) individual customer welfare CW^* weakly increases;

(ii) if the average search reward ER^* decreases (ER^* under s is lower than $ER^{*\prime}$ under s', where $s < s'$), there must exist an intermediate search cost $\tilde{s} \in (s, s')$ under which customers randomize.

Proposition 2.4-(i) shows that decreasing search costs will not hurt customer welfare. This result holds regardless of whether the underlying equilibrium consists of pure or mixed strategies. However, it does not rule out the possibility of customer welfare being unchanged under mixed strategy equilibria as shown in Proposition 2.2. The appearance of mixed strategy equilibria is a critical characteristic of quality differentiation, and Proposition 2.4-(ii) suggests that it is also a paramount prerequisite for lower search costs to degrade search reward. This necessary condition underscores the significance of quality differentiation (and the concomitant quality substitution effect). From Example 2.1, mixed strategy equilibria tend to arise at small search cost levels. Therefore, Proposition 2.4-(ii), to a certain extent, identifies significant quality differentiation and small search costs as the main drivers for the non-monotone comparative statics. Recall from Theorem 2.2 that Pareto dominant mixed strategy equilibria do not appear if servers are identical in quality ($N = 1$). An intuitive corollary immediately follows that if $N = 1$, reducing search costs always improves the average search reward, which is equivalent to lowering waiting time since quality is identical across servers. Therefore, quality differentiation ($N \geq 2$) is essential for our results to hold.

2.6 Impact of Policy Intervention 2: Arrival Rate Reduction

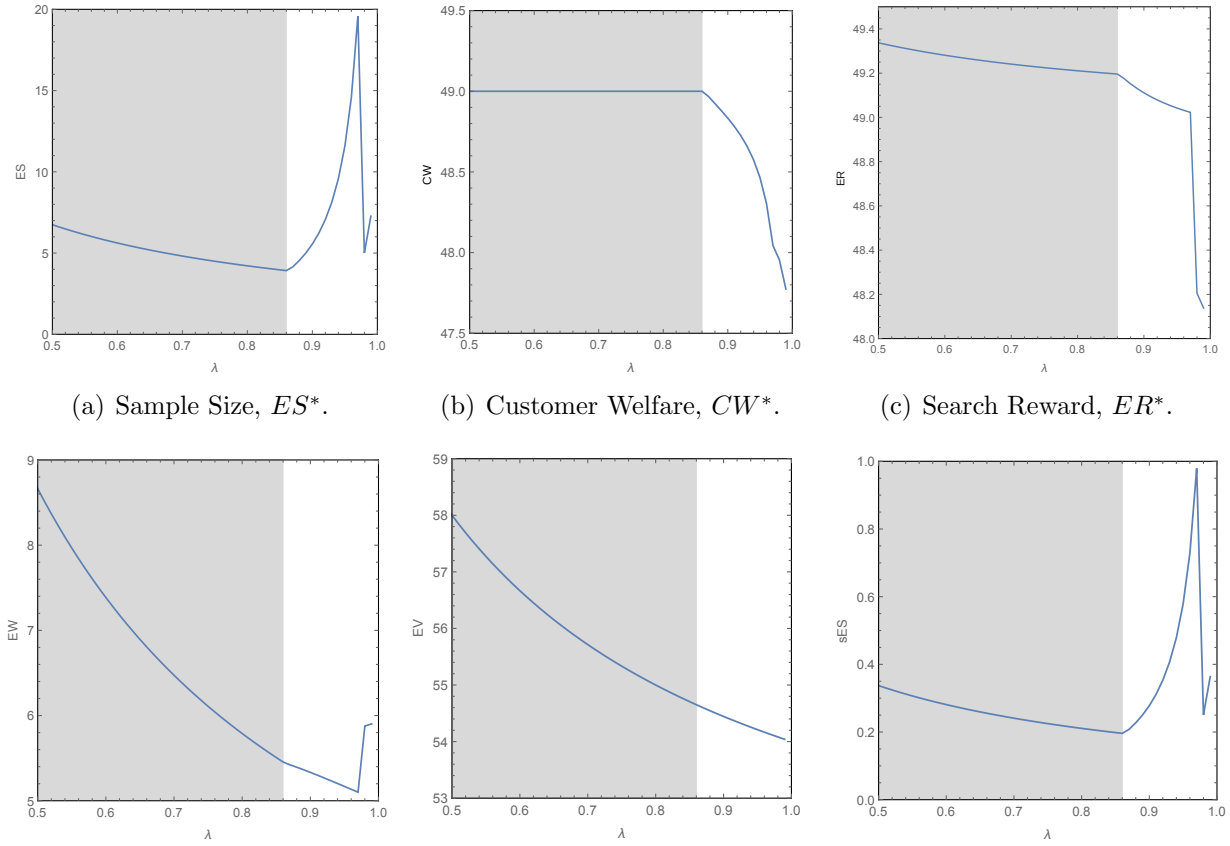
In this section, we analyze how a change in the arrival rate impacts customer and system behavior. As the arrival rate falls, a lower system load puts a downward pressure on the average waiting time. Unlike the change in search costs, which only makes an impact on the average waiting time when the search thresholds are changed, the change in the arrival rate can influence the average waiting time even when the search thresholds remain the

same. Hence, the implications of reducing the arrival rate are much more subtle than the implication of reducing search costs.

2.6.1 An Illustrative Example

Example 2.2. *Consider the following parameters:* $N = 2$, $\mu = 1$, $s = 0.05$, $c = 1$, $\boldsymbol{p} = (0.6, 0.4)$, $\boldsymbol{V} = (50, 60)$. *The system metrics versus the arrival rate are plotted in Figure 2.2.*

Figure 2.2: System metrics versus arrival rate for Example 2.2.



(a) Sample Size, ES^*. (b) Customer Welfare, CW^*. (c) Search Reward, ER^*.

(d) Waiting Time, EW^*. (e) Quality, EV^*. (f) Search Expenditure, sES^*.

Note. Mixed-strategy equilibria are indicated by shaded areas.

Figure 2.2-(a) illustrates that lowering the arrival rate leads to less sampling under pure strategy equilibria and yet more sampling under mixed strategy equilibria. Figure 2.2-(b)

illustrates that individual customer welfare might fail to strictly increase as the arrival rate falls, an violation of Principle 4(c). Figure 2.2-(c) shows that the average search reward tends to increase with a reduction in the arrival rate. Figure 2.2-(d) indicates that the average waiting time may increase with a lower arrival rate, which directly conflicts with Principle 4(a). Moreover, it is also at odds with Principle 5(a) in the sense that one would expect a lower arrival rate (reducing the load) and sampling more under mixed strategy equilibria (balancing the load) to decrease the average waiting time; nevertheless, they jointly engender a longer wait. Figure 2.2-(e) illustrates that customers may obtain higher quality as the arrival rate falls. In Figure 2.2-(f), the search expenditure changes in an identical way as the amount of sampling in Figure 2.2-(a) since search cost s is unchanged.

In summary, reducing the arrival rate may not improve individual customer welfare, and may strictly increase the average waiting time. These numerical findings are in disagreement with Principles 4 and 5. As we shall see with our analytical development, some of these numerical findings are tied to the chosen model primitives while others are more general.

2.6.2 Analytic Results

Proposition 2.5 below identifies conditions for a lower arrival rate to lead to more sampling, a longer waiting time, and unimproved individual customer welfare, violating Principles 4 and 5. The underlying equilibrium structure is similar to that in Proposition 2.2: customers are indifferent (and thus randomize) between searching and joining an empty server of the lowest quality level. It corresponds to the shaded area in Figure 2.2.

Proposition 2.5. *Consider $N \geq 2$. If search cost s is small enough, then under an intermediate arrival rate $\lambda \in (\underline{\lambda}, \overline{\lambda})$, where $\overline{\lambda}, \overline{\lambda} - \underline{\lambda}$ are both decreasing in s, customers play a mixed strategy equilibrium with search threshold $k_1^* \in (0,1)$ for the lowest quality level and $k_i^* = 1 + \lfloor d_i \rfloor$ for $i = 2, \ldots, N$. As arrival rate λ decreases within $(\underline{\lambda}, \overline{\lambda})$:*

(i) customers sample more servers before joining, i.e., ES^* increases;

(ii) customer welfare CW^* remains unchanged;

(iii) customers obtain higher quality on average, i.e., EV^* increases;
(iv) the average waiting time EW^* increases if $\lfloor d_N \rfloor \geq 1$ and remains unchanged otherwise;
(v) the average search reward ER^* increases.

Under the conditions specified in Proposition 2.5, Principle 4(c) is violated; due to randomization, individual customer welfare is unaffected when the arrival rate is reduced. This case, however, does obey Principle 4(b). Since $ER^* = sES^* + CW^*$, more sampling and constant individual customer welfare immediately imply that the search reward ER^* improves (Proposition 2.5-(v)). Due to the quality substitution effect shown in Proposition 2.5-(iii), reducing the arrival rate actually leads to an increase in the average waiting time (Proposition 2.5-(iv)), violating Principle 4(a). Here, like the case in Proposition 2.2, customers sample more in the direction of screening out more servers of the lowest quality level and crowding at higher quality servers with longer queues. More sampling is not conducive to load balancing, but, rather, imposes an upward pressure on the average waiting time. What about the downward pressure from a lower system load? In this case, the decrease in demand rates is completely offset by the increase in the amount of sampling such that the rate at which customers inspect each queue, search intensity $\alpha(\boldsymbol{k})$, remains constant (see Equation 2.2). Hence, the quality substitution effect becomes the only force to drive up the average waiting time.

The key condition in Proposition 2.5 is small search costs, reminiscent of the environment identified in §2.5. Moreover, the interval $(\underline{\lambda}, \bar{\lambda})$ gets wider and shifts up when the search cost is smaller. This further demonstrates that small search cost levels make the phenomena specified in Proposition 2.5 more pervasive and emerge when the system load is high (presumably more relevant for practice). Finally, we show in the appendix that $\bar{\lambda}$ is increasing in integer d_N, which implies that this case will be most relevant in a system with greater quality differentiation. Again, much in the same spirit of §2.5, this points to the new challenge this policy intervention has to confront when search frictions become low with technological advances and servers are noticeably different in quality.

Complementing Proposition 2.5, which pertains to particular settings under mixed strategy equilibria, we now characterize, more generally, the driving forces underlying the impact of reducing the arrival rate under pure strategy equilibria. For any pure strategy equilibrium under arrival rate λ, changing the arrival rate in a small neighborhood around λ will not change the equilibrium search thresholds. This follows from the recognition that $CW(\boldsymbol{k}; \boldsymbol{\pi}(\boldsymbol{k}))$ is continuously differentiable in λ for any given $\boldsymbol{k}$. Thus, it is well posed to consider the impact of a small change in the arrival rate on the system metrics to the extent of unchanged search thresholds. From a practical standpoint, starting our analysis with a small change in the arrival rate is an important first step given that diverting demand from public health systems is likely to be executed in a piecemeal fashion. In Proposition 2.6, by saying the arrival rate decreases under a pure strategy equilibrium, we mean that the search thresholds are unchanged with the arrival rate reduction.

Proposition 2.6. *Consider a pure strategy equilibrium* $\boldsymbol{k}^*$.
Define $\hat{\lambda} = \mu \left[1 - \sum_{i=1}^{N} p_i (k_i^* + 1)^{-1}\right]$. *When arrival rate* λ *decreases under the pure strategy equilibrium* $\boldsymbol{k}^*$,

(i) customers sample less, i.e., ES^* *decreases;*
for $N \geq 2$,

(ii-a) when the arrival rate is low, decreasing in the range $(0, \hat{\lambda})$, *both the average quality obtained* EV^* *and the average waiting time* EW^* *decreases;*

(ii-b) when the arrival rate is high, decreasing in the range $(\hat{\lambda}, \mu)$, *the average quality obtained* EV^* *increases and the average waiting time* EW^* *may either increase or decrease.*

Proposition 2.6-(i) is interesting. While reducing the search cost always makes customers sample more (as search theory would predict, see Principle 3), reducing the arrival rate has an ambiguous impact on customer search behavior. Even when the search thresholds are fixed with the arrival rate reduction, the queue length distribution changes. Proposition 2.6-(i) specifically shows that queue length distribution shifts in a way that customers are less

likely to encounter long queues they would reject, and therefore *less* sampling is entailed, in contrast to Proposition 2.5-(i). This analytically confirms the numerical observation of the non-monotone sampling behavior in Figure 2.2-(a). From an intuitive perspective, a lower system load means less competition among customers, which may suggest that they would obtain higher quality on average; on the other hand, customers sample less, and, thus, may be more prone to joining low quality servers.

Proposition 2.6-(ii) indicates that customers may, indeed, substitute toward either high or low quality servers, subtly depending on how the arrival rate compares to the capacity (given the search thresholds). The turning point is $\hat{\lambda}$, which is determined by setting search intensity $\alpha(\boldsymbol{k}^*) = 1$. Proposition 2.6-(ii-a) corresponds to the case in which customers obtain lower quality service on average, which arises when the arrival rate is relatively small (the resulting search intensity $\alpha(\boldsymbol{k}^*)$ being less than 1), which implies competition among customers is not tense to begin with, and, therefore, the quality substitution effect is mostly driven by less search (hence lower quality obtained), and, therefore, works in the same direction as lower system load, and the average waiting time in the system declines.

Proposition 2.6-(ii-b) shows that the alternative case arises when the arrival rate is relatively high (the resulting search intensity $\alpha(\boldsymbol{k}^*)$ being greater than 1). In this case, customers are competing fiercely, and, therefore, with a lower system load, the quality substitution effect is mostly driven by less competition, which implies customers would obtain higher quality of service on average. Since customers "flock" to high quality servers that attract longer queues, the quality substitution effect puts an upward pressure on the average waiting time in the system. Unlike the case identified in Proposition 2.5, now the downward pressure from lower system load is no longer silent because the decrease in both the arrival rate and the amount of sampling implies a decline in search intensity $\alpha(\boldsymbol{k})$ (see Equation (2.2)). Thus, the overall impact becomes nebulous. If the upward pressure from quality substitution outweighs the downward pressure from lower system load, then the average waiting time will get longer, violating Principle 4(a). This is indeed possible, as illustrated by the longer waiting time

under pure strategy equilibria in Figure 2.2-(d) and the associated higher quality obtained in Figure 2.2-(e) (as the arrival rate falls).

Note that $\hat{\lambda}$ is increasing in k_i for all i. Recall from Proposition 2.3 that a lower search cost would lead to smaller search thresholds. Putting these together implies that the range $(\hat{\lambda}, \mu)$ gets wider at smaller search cost levels. Thus, the scenario described in Proposition 2.6-(ii-b) is more likely to be encountered when the arrival rate is large, characteristic of high demand service systems, and when search costs are small, an increasingly relevant case with improved technologies.

Proposition 2.6 deals with the case in which the reduction in the arrival rate does not change the search thresholds; in Appendix B.5, we provide analytical results that identify conditions under which reducing the arrival rate or both the search cost and arrival rate *shift* down the pure strategy search thresholds in equilibrium (load balancing) and still increase the average waiting time. This represents a case in which the upward pressure from the quality substitution effect overshadows the downward pressure from both the direct effect of a lower system load and the indirect effect of load balancing. Our main findings are (1) quality differentiation should be substantial; (2) search cost should be low such that the induce equilibrium search thresholds are small (cf. Proposition 2.3-(i)). These findings are consistent with our results in Proposition 2.5 and Proposition 2.6.

Up to this point, our discussion is mute on how the search reward fares as a result of arrival rate reduction. In Example 2.2, the negative consequence of a longer wait is overshadowed by the positive improvement in obtained quality, so the average search reward always changes for the better (Figure 2.2-(c)). However, this result does not hold in general, and we present such a counter-example in Appendix B.4. We choose not to highlight this result because our numerical studies suggest that this tends to happen when the arrival rate is very small relative to the capacity (corresponding to the conditions identified in Proposition 2.6-(ii-a)), and hence may not be vastly relevant.

Given the complicated dependency of the equilibrium conditions on the arrival rate λ,

proving average customer welfare is monotonically (weakly) decreasing in the arrival rate turns out to be analytically intractable. However, we do not find a counter-example that numerically defies this conjecture. Still, from a queueing-theoretic perspective, we believe finding that the average waiting time can be non-monotone in the arrival rate and characterizing conditions under which this occurs are interesting in their own rights.

2.7 Conclusion, Policy Implications, and Discussion

The present chapter captures the three-way trade-off faced by customers when they decide which queue to join: heterogeneity in servers due to vertical differentiation, waiting costs due to congestion, and search frictions due to the lack of full information. Our results show that a reduction in either the search cost or arrival rate may not improve individual customer welfare and lead to a strict increase in the equilibrium average waiting time in the system as customers strategically substitute toward high quality servers that entail a longer wait. Moreover, as search costs fall, the higher quality customers obtain may not offset the longer wait, and the search reward deteriorates. We find these phenomena tend to arise under relatively large quality differentiation and relatively small search cost levels. In Appendix B.4, we conduct additional numerical experiments that confirm the robustness of our findings.

Our results have practical implications for public health systems where patients seeking elective surgeries are subject to long wait times. It is important to recognize that policy interventions either to reduce search frictions (such as launching quality and wait times websites) or to dampen demand (such as introducing privately funded health care) may create distorted incentives, yielding no improvement in customer welfare, lengthening wait times and in the case of reducing search frictions, worsening the reward patients get from service providers. Setting up these policy interventions usually involves a variety of non-negligible implementation costs, such as establishing the infrastructure to accurately gather data on wait times and safety records, subsidizing private care to support its presence, and

adjusting the budget allocated to public systems. If there are zero returns (sometime even negative returns in terms of wait times) to these investments, then the interventions may not be well justified. Of course, the extent to which the effects identified in this chapter play a role in practice is an empirical question. Thus, to comprehensively assess the impact of these policies, health authorities should pay close attention to the environment in which they are operating, particularly, if there is significant quality disparity among healthcare providers, and if search frictions are already relatively small, both of which are key contributors to the aforementioned distortions. As the boom in information technology tremendously facilitates patient choice (low search frictions), this presents a new challenge for health authorities.

Our results about how policies may backfire in addressing long wait times are instrumental on a practical level. As compared to service providers' quality, wait times are more easily measurable, more likely to be a source of public outcry, and, therefore, more often used as an explicit policy target. We contribute to this ongoing policy debate from an alternative perspective.

With regard to the first policy proposal (decreasing search cost s), Siciliani and Hurst (2005) argue that it may help reduce waiting time disparities by directing patients from high demand areas to low demand areas. This argument overlooks the fact that this policy intervention may create an incentive among patients to focus more on the quality dimension and thus substitute toward higher quality surgeons with longer waits. Evidence suggests that when provided with choices, patients are likely to base their decision on factors like reputation and doctor's opinion, more so than wait times (see Fotaki et al. 2008, Thomson and Dixon 2006). Our model provides some theoretical justifications for these findings. Even worse, our model indicates that the improvement in quality obtained may not justify the increase in wait times, thereby worsening the search reward.

As for the second policy proposal (decreasing arrival rate λ), advocates typically postulate the "demand-side effect": as patients switch to the private system, less demand for the public wait list reduces its average waiting time (Chen et al. 2015). Much of the opposition to private

clinics typically appeals to the "supply-side effect," whereby private systems crowd out the supply of surgeons to public systems (e.g., Rachlis 2005), thus inducing a longer wait. We show that even when supply is fixed, the "demand-side effect" still does not necessarily favor privately funded health care when customers' strategic search behavior is taken into account. In the presence of quality differentiation, patients may crowd toward high quality surgeons when faced with lower demand, driving up wait times in the public system.

While wait times is a more convenient metric due to its high visibility, we note that there are times when customer welfare and/or the search reward improve despite longer wait times. Thus, if policymakers' goal is primarily enhancing customer welfare, then using wait times as a surrogate for measuring whether this goal has been attained may lead to a biased evaluation.

2.7.1 Discussion and Future Research

Future research could incorporate more details in modeling customer search behavior. For example, real-world customers may not necessarily search sequentially, but rather combine sequential search and simultaneous search (Morgan and Manning 1985). Moreover, our model assumes that the search for quality and waiting time is performed concurrently, whereas, in reality, customers may engage in two separate search processes, one on quality and one on waiting time, given the distinct characteristics of these two dimensions. This would present modeling challenges if customers search along the queue length dimension first and infer servers' quality from their observed queue length. A model the combines herding and search would be entailed. A related theoretical question is how customers determine their selection rule (Weitzman 1979) if all queues are observable but costly search is required to resolve uncertainty about quality. These subtle aspects of customer search behavior can be directions for future research.

We assume customers to be identical in our model, although there might be significant heterogeneity in the customer population in practice. Notably, they can be different in

(1) search costs, (2) waiting costs, and (3) their valuation of different servers. (1) We can accommodate this extension by finding the search thresholds for each type of customers: customers with higher search costs are less selective and thus adopt higher search thresholds. (2) Heterogeneity in waiting costs brings up the issue of prioritization in queues. This is also a common practice in health care as surgeons might prioritize their treatment of patients based on severity. (3) We assume that servers are vertically differentiated, but there could also be an element of horizontal differentiation where customers differ in their rankings of servers. In the health care context, one source for horizontal differentiation is geographical proximity. On a different note, service rates in our model are common knowledge and identical across servers, whereas it is possible that service rates may be correlated with service value (quality). Debo and Veeraraghavan (2014) show that queueing joining behavior in such an environment has "sputtering equilibria."

Finally, our work only examines the strategic response of customers (the demand side) to policy changes, and turns off the strategic aspect of service providers (the supply side). Service providers, in general, may respond by changing capacity, prices, or quality provided; they may aslo be motivated to cherry-pick customers. If a parallel private system is introduced that drains the supply to the public system as surgeons switch, then this "supply-side effect" only strengthens our results on longer waits. Reducing search frictions may intensify competition among service providers in capacity investment, which may be beneficial to waiting time reduction. However, as Chen et al. (2016) show, service providers may also be subject to a signaling effect, whereby high quality hospitals may choose low capacity and use long waits to lure patients uninformed about quality. Queues as a signaling device have received a growing research interest (e.g., Debo et al. 2012, Veeraraghavan and Debo 2011). When service providers are allowed to adjust their prices, the quality substitution effect may not be as pronounced as in our current model since high quality servers have an incentive to charge a higher price and restrict throughput. On the other hand, Dranove and Satterthwaite (1992) show that improved price information may decrease welfare as intensified price

competition makes firms select severely suboptimal level of quality. Dranove et al. (2003) find evidence suggesting that healthcare report cards may decrease welfare, as hospitals are incentivized to decline more difficult, severely ill patients. Our model is a first step in understanding customer search behavior in a congested environment with quality differentiation, and encompassing the supply side behavior warrants further research.

CHAPTER 3
INVITE YOUR FRIEND AND YOU'LL MOVE UP IN LINE: LEVERAGING SOCIAL TIES VIA OPERATIONAL INCENTIVES

3.1 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' dis-

like 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 chapter 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, re-

ferred 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 chapter is organized as follows. §3.2 reviews related literature. In §3.3, we set up the main model of the joining-referral game among customers in the referral priority program. §3.4 characterizes the equilibrium of the game. In §3.5, we examine when customers would refer, and when the referral priority program would be (in)effective in improving system throughput and customer welfare. §3.6 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. §3.8 concludes the chapter with a summary of our main results, and discusses future research.

3.2 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.

3.2.1 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.

The present chapter 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. This chapter 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.

3.2.2 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. The present chapter 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. This chapter 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, this chapter 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.

3.3 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:

$$U_{\sigma',\sigma}(v,\chi) = \begin{cases} 0, & (j', r') = (0,0) \\ v - cW_2^{\chi}(\sigma), & (j', r') = (1,0) \\ v - c_r - c[\alpha(\sigma)W_1^{\chi}(\sigma) + (1-\alpha(\sigma))W_2^{\chi}(\sigma)], & (j', r') = (1,1) \end{cases} . \tag{3.1}$$

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, $\chi \in \{B, R\}$.

3.3.1 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 §3.4.1). 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 §3.4.1) 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 chapter.

3.4 Equilibrium

In this section, we characterize the equilibria of the referral-joining game. First, in §3.4.1, 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 §3.4.2 using the expressions derived in §3.4.1.

3.4.1 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:

$$\mathbb{P}(N=1) = 1-q, \quad \mathbb{P}(N=k) = q(1-p)p^{k-2}, k \geq 2, \qquad \mathbb{E}[N] = \frac{1+q-p}{1-p}, \tag{3.2}$$

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. 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 (3.2); 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 3.1 gives closed-form expressions of these expected delays.

Lemma 3.1. *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 3.1 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 3.1 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 3.1.

Assumption 3.1. $\bar{V} > 2c/\mu$.

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

3.4.2 Equilibrium Referral Strategies

Equipped with the expected delay expressions in Lemma 3.1 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 (3.1), 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

$$c_r \leq c\alpha(\sigma)[W_2^{\chi}(\sigma) - W_1^{\chi}(\sigma)]. \tag{3.3}$$

The left-hand side (LHS) of (3.3) is the cost of referring a friend, and the right-hand side (RHS) of (3.3) 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 3.1. *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 3.1 directly follows from Lemma 3.1 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 (3.3), it follows that referred customers have a greater incentive to refer against any equilibrium conjecture. The intuition is as follows. From Lemma 3.1, 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.

Strict Non-referral Equilibrium: $(r^B, r^R) = (0, 0)$.

Neither base nor referred customers refer. The equilibrium (β, α) satisfies the following conditions:

$$\bar{V}(1-\beta) - \frac{c}{\mu - \Lambda\beta} = 0, \tag{3.4a}$$

$$c_r \geq c\alpha \left(\frac{1}{\mu - \Lambda\beta} - \frac{1}{\mu} \right), \tag{3.4b}$$

$$\text{If } \bar{V} - c\left[\frac{1}{\mu - \Lambda\beta} + \frac{1}{\mu}\right] < 0, \ \alpha = 0; \quad \text{otherwise, } \bar{V}(1-\alpha) - c\left[\frac{1}{\mu - \Lambda\beta} + \frac{1}{\mu}\right] = 0. \tag{3.4c}$$

This strict non-referral equilibrium results in an $M/M/1$ queue with expected delay $1/(\mu - \Lambda\beta)$ and throughput $\Lambda\beta$. Equation (3.4a) 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 (3.4b) 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 (3.4b) 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 3.1, (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 (3.4c) determines α in much the same way as equation (3.4a): $\bar{V}(1-\alpha) = v^R$ is the marginal referred customer's valuation for service.

Because the LHS of equation (3.4a) is decreasing in β, we can uniquely pin down β. Plugging this β into Condition (3.4c) uniquely determines α. Thus, (3.4b) 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.

All-Referral Equilibrium: $(r^B, r^R) = (1, 1)$.

Both base and referred customers refer. The equilibrium (β, α) satisfies the following conditions:

$$\bar{V}(1-\beta) - c_r - c[\alpha W_1^B(\alpha,\beta) + (1-\alpha)W_2^B(\alpha,\beta)] = 0, \tag{3.5a}$$

$$c_r \leq c\alpha \left[W_2^B(\alpha,\beta) - W_1^B(\alpha,\beta)\right], \tag{3.5b}$$

$$\bar{V}(1-\alpha) - c_r - c[\alpha W_1^R(\alpha,\beta) + (1-\alpha)W_2^R(\alpha,\beta)] = 0, \tag{3.5c}$$

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 (3.5a) and (3.5c) jointly determine α and β by requiring that marginal base and referred customers have zero expected utility, assuming they refer. Condition (3.5b) 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.

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:

$$\bar{V}(1-\beta) - c/(\mu - \Lambda\beta) = 0, \tag{3.6a}$$

$$c_r = c\alpha\left(\frac{1}{\mu-\Lambda\beta} - \frac{1}{\mu}\right), \tag{3.6b}$$

$$\bar{V}(1-\alpha) - c\left[\frac{1}{\mu-\Lambda\beta} + \frac{1}{\mu(1-\kappa\alpha)}\right] = 0. \tag{3.6c}$$

In this weak non-referral equilibrium, the throughput is $\Lambda\beta$, where β is determined by equation (3.6a), which is the same as equation (3.4a), 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 (3.6b) and (3.6c). Equation (3.6b) ensures (fictitious) referred customers are indifferent to referrals, which supports their mixed referral strategy. In fact, by Corollary (3.1), 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 (3.6b) by plugging β from equation (3.6a). Condition (3.6c) 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 (3.6c) 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).

Partial-Referral Equilibrium: $(r^B, r^R) = (\kappa, 1)$.

Base customers randomize; referred customers refer. The equilibrium (β, α, κ) satisfies the following conditions:

$$\bar{V}(1-\beta) - c_r - c[\alpha W_1^B(\alpha,\beta,\kappa) + (1-\alpha)W_2^B(\alpha,\beta,\kappa)] = 0, \tag{3.7a}$$

$$c_r = c\alpha[W_2^B(\alpha,\beta,\kappa) - W_1^B(\alpha,\beta,\kappa)], \tag{3.7b}$$

$$\bar{V}(1-\alpha) - c_r - c[\alpha W_1^R(\alpha,\beta,\kappa) + (1-\alpha)W_2^R(\alpha,\beta,\kappa)] = 0, \tag{3.7c}$$

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 (3.7a) through (3.7c). Conditions (3.7a) and (3.7c) require that marginal base and referred customers have zero expected utility if they choose to refer. Condition (3.7b) 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.

3.4.3 Existence of Equilibria and Structural Results

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

$$\bar{\Lambda} \equiv \mu \frac{\bar{V}(\bar{V} - 2c/\mu)}{(\bar{V} - c/\mu)c/\mu}. \tag{3.8}$$

One can show by straightforward algebra that $\bar{\Lambda}$ is increasing in $\bar{V}$. Building on the definition of $\bar{\Lambda}$, Proposition 3.1 establishes structural properties of the all-referral equilibrium.

Proposition 3.1. *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 3.1 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 (3.5a) and (3.5c). 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 3.1, we formally establish the existence of equilibria in Theorem 3.1.

Theorem 3.1. *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* $c_r \in \Big[\min\{c_r^l, c_r^m\}, \max\{c_r^l, c_r^m\}\Big]$.

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 3.1 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 3.1, 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 3.1 and Theorem 3.1, 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 3.2. *If the referral cost is zero* ($c_r = 0$)*, there exists a unique pure strategy equilibrium.*

Corollary 3.2 follows from Proposition 3.1 and Theorem 3.1. 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 3.3. *In any equilibrium, the joining probability of referred customers is lower than that of base ones ($\alpha < \beta$).*

Corollary 3.3 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 3.1). 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.

3.5 Effectiveness of the Referral Priority Program

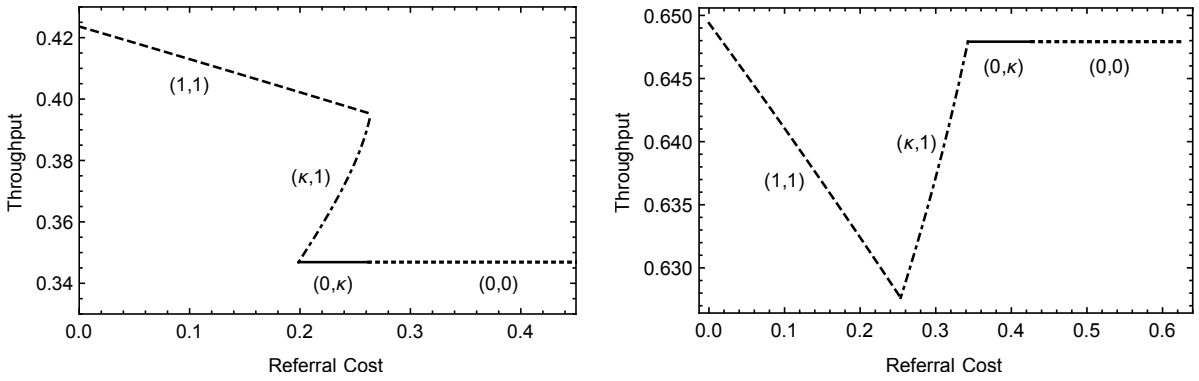
The focus of this section is to study when the referral priority program is effective in acquiring more customers/increasing throughput as a marketing tool. Specifically, we examine when customers would refer, when referrals would have a positive effect on throughput, and when it may have unintended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

3.5.1 Two Illustrative Examples

Figure 3.1 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 3.1. Third, when $\Lambda = 0.5$ (Figure 3.1-(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 3.1-(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.

Figure 3.1: 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.

3.5.2 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 3.2. *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* $\bar{V}$ *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 3.2 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 (3.8)) persists. This result ties back to our discussion of Proposition 3.1: a base market size that is too large prevents referred customers from joining.

Moreover, Proposition 3.2 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 3.3. *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 3.3 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 3.2. *The referral priority program induces an equilibrium that reduces the throughput relative to FIFO ($\lambda^R < \lambda^F$) only if $\Lambda > \mu/2$.*

Theorem 3.2 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 3.2 analytically confirms the observation from Figure 3.1-(a) that the referral priority program would never reduce the throughput when $\Lambda = 0.5$ and $\mu = 1$.

Theorem 3.3. *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 3.3 complements Theorem 3.2 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 3.1 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 3.3 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 3.1-(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.

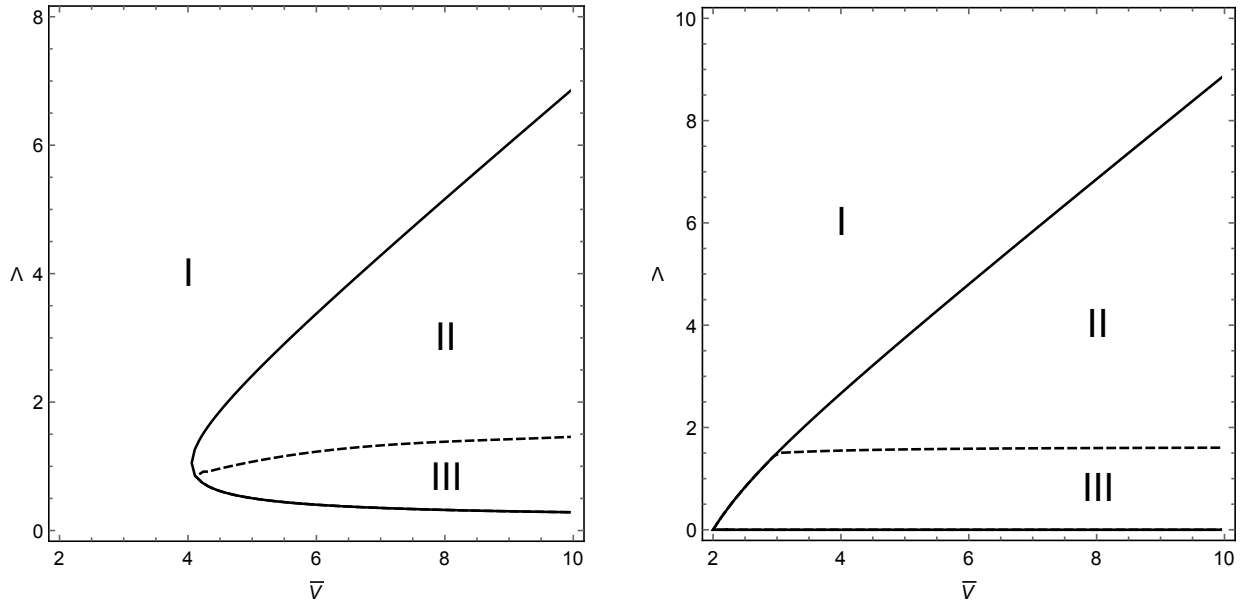
Proposition 3.4. *For any $\bar{V} > 3c/\mu$, there exists a sufficiently small $\epsilon > 0$ such that under $\bar{\Lambda} - \epsilon$ and $c_r = 0$, the referral priority program reduces the throughput relative to FIFO ($\lambda^R < \lambda^F$).*

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 3.1 suggests the all-referral equilibrium achieves the highest throughput when $c_r = 0$. However, Proposition 3.4 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 3.1). 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 3.2 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 3.2: Effectiveness of the referral priority program.



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

Note. 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. $c = 1,\ \mu = 1$. With $c_r = 0$ in panel (b), equilibrium uniqueness is guaranteed due to Corollary 3.2. 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 3.2-(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 3.2-(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 3.2-(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 3.2-(b) also illustrates Proposition 3.4: 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 3.5. *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 3.5 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 3.3) and weaker joining incentives of referred customers (Corollary 3.3). 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.

3.6 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. §3.6.1 examines the optimal pricing problem in the referral priority program. In §3.6.2, we introduce the referral reward program and lay out the firm's joint optimization problem of the admission price and referral reward. §3.6.3 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 (3.9a)-(3.9b):

$$\max_{P\geq 0,\ \beta\in[0,1]} \quad P\Lambda\beta \tag{3.9a}$$

$$\text{s.t.} \quad \bar{V}(1-\beta) - P - \frac{c}{\mu - \Lambda\beta} = 0, \tag{3.9b}$$

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

3.6.1 Pricing in the Referral Priority Program

Recall from §3.4.2 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 3.2). 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 C.2. We denote the referral priority program with optimal pricing by the optimal referral priority program.

Proposition 3.6. *The optimal referral priority program generates referrals only when the base market size Λ is intermediate.*

Proposition 3.6 is an extension of Proposition 3.2. 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.

3.6.2 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 (3.9a)-(3.9b) 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 3.2). 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 3.2. *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

$$\max_{P\geq 0, \Delta \geq 0;\ (\alpha,\beta,\kappa)\in[0,1]^3} \quad P\Lambda\beta + (P-\Delta)\frac{\Lambda\beta\kappa\alpha}{1-\kappa\alpha} \tag{3.10a}$$

$$\text{s.t.} \quad \bar{V}(1-\beta) - P - cW^B(\alpha, \beta, \kappa) + \kappa(-c_r + \alpha\Delta) = 0, \tag{3.10b}$$

$$\bar{V}(1-\alpha) - P - cW^R(\alpha, \beta, \kappa) + \kappa(-c_r + \alpha\Delta) = 0, \tag{3.10c}$$

$$\text{either} \ \ \kappa = 0, c_r \geq \alpha\Delta, \ \ \text{or} \ \ \kappa \in (0,1), c_r = \alpha\Delta \ \ \text{or} \ \ \kappa = 1, c_r \leq \alpha\Delta. \tag{3.10d}$$

In objective function (3.10a), 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 (3.10b) and (3.10c) are participation constraints for base and referred customers, respectively. Conditions (3.10d) 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 (3.10a)-(3.10d) to (3.11a)-(3.11c) 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.

$$\max_{P^e \geq 0;\ (\alpha,\beta,\kappa)\in[0,1]^3} \quad P^e \frac{\Lambda\beta}{1-\kappa\alpha} \tag{3.11a}$$

$$\text{s.t.} \qquad \bar{V}(1-\beta) - P^e - cW^B(\alpha,\beta,\kappa) - \kappa c_r = 0, \tag{3.11b}$$

$$\bar{V}(1-\alpha) - P^e - cW^R(\alpha,\beta,\kappa) - \kappa c_r = 0. \tag{3.11c}$$

If the optimal solution to (3.11a)-(3.11c) 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 3.7. *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 3.6), 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.

3.6.3 Numerical Comparison

We conduct a numerical study to compare the profit performance of the two referral programs. Table 3.1 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 3.1 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 3.1.

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 3.6 and 3.7. 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' in-

Table 3.1: 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)

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

centive 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 3.2. 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 3.6 and 3.7, 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 §3.5 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 3.2). 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 C.2, 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.

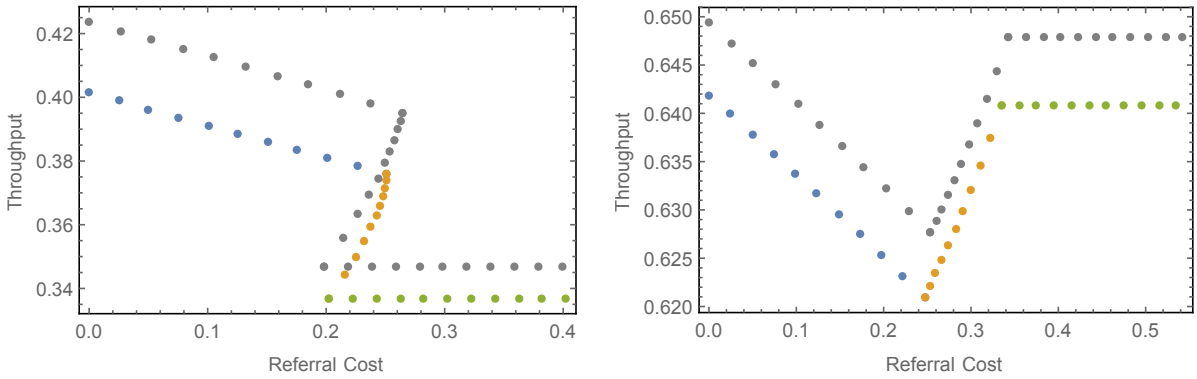
3.7 Extensions

3.7.1 Varying the Service Rate

Proposition 3.2 indicates that when the base market size Λ is small relative to the service rate μ, customers do not refer. The begs the question of whether, for a fixed base market size, decreasing the service rate would motivate customers to refer, thereby improving the system throughput. If so, the service provider would have an incentive to slow down since decreasing the service rate is usually easy to implement whereas increasing the service rate entails capacity investment. Thus, we numerically explore how varying the service rate impacts customers' referral behavior. We illustrate our findings with Figure 3.3. It plots throughputs against referral costs as in Figure 3.1 under two distinct service rates, $\mu = 1$ and $\mu = 0.95$.

In Figure 3.3-(a) (when $\Lambda = 0.5$), for some referral cost c_r (0.22-0.25), there exists a non-referral equilibrium under high service rate ($\mu = 1$) and a referral-equilibrium under low service rate ($\mu = 0.95$); and the throughput of the former equilibrium could be lower than

Figure 3.3: Throughput against referral cost c_r under different service rate μ.



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

Note. $\bar{V} = 5$, $c = 1$, $\mu = 1$ for the gray dots; $\mu = 0.95$ for the colored dots.

that of the latter. This observation seems to corroborate the aforementioned conjecture that the service provider may want to intentionally slow down (limit supply) to generate higher demand. However, for those c_r's, there exist *multiple* equilibria under both service rates. While we can find an equilibrium under a low service rate that achieves a higher throughput than one under a high service rate, we can also find a different equilibrium under a service rate that throughput-dominates the above two equilibria. Thus, because of multiplicity of equilibria, it is unfair to cherry-pick equilibria to favor one service rate over another in comparing throughputs under different service rates. While we do not attempt equilibria refinement to pin down a unique equilibrium, it is still useful to compare how throughputs of a set of equilibria shift when the service rate is varied. For instance, for a given set of parameters, we can track two equilibria, one that achieves the maximum throughput (a) and one that achieves the lowest throughput (b). Figure 3.3-(a) suggests that both (a) and (b) under the high service rate ($\mu = 1$) throughput-dominates their counterparts under the low service rate ($\mu = 0.95$). Thus, it is generally hard to argue that reducing the service rate would definitively encourage referrals benefit the system throughput.

Figure 3.3-(b) (when $\Lambda = 1.5$) shows a case in which equilibria are unique and the throughput under high service rate $\mu = 0.95$ is consistently higher than the throughput

under low service rate ($\mu = 0.95$) across different referral costs. Moreover, when the service rate is smaller, the threshold referral cost for customers to refer is smaller, suggesting a bigger referral reluctance from customers. This is because customers expect it less likely for their referrals to convert, i.e., their friend is less likely to join due to high congestion caused by low service rate. In this example, again, reducing the service rate does not prompt customers to refer more and does not boost the system throughput.

3.7.2 Challenges with Observable Queues

One 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. 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. Incorporating such dynamic behavior is complicated even in the setting of Afèche and Sarhangian (2015), in which referrals are not considered and priority is pre-assigned rather than self-selected.

If we restrict attention to referrals upon arrivals, while a similar problem is well studied for priority purchasing (such as Adiri and Yechiali 1974, Hassin and Haviv 1997, Cui et al. 2017), referrals introduces specific challenges and 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). This is because a long queue would deter referred customers from joining, and thus cause existing customers not to refer. We provide some formalization of this idea in the following.

We consider an observable queueing model with a finite buffer size N. One can interpret the buffer size as the Naor threshold $N = \lfloor V\mu/c \rfloor$ assuming customers do not know the referral program when they make joining decisions and that all customers have a common

value for service V). Customers decide whether to refer upon arrival based on the queue length they observe. They assume the size of both the priority class and the regular class. As in the unobservable model, we assume that referral response is instantaneous. We further assume that a referred friend joins with probability p, exogenously specified.

It is established in the literature of priority purchasing (e.g., Hassin and Haviv 1997) that customers purchase priority if the total queue length upon arrival is weakly above a certain threshold n. Since customers are ex-ante symmetric, customers can perfectly infer the size of each priority class by only observing the total queue length. Specifically, if the total queue length is less than or equal to n, then all the customers in the queue are in the regular class; if the total queue length is greater than n, then exactly n customers in the queue are in the regular class and the rest of customers (at the head of the queue) are in the priority class. However, this result of threshold strategies does not carry over easily to the referral priority program.

Define a threshold referral strategy to be making a referral if and only if the total queue length upon arrival is at least n (a customer seeing queue length $N-1$ does not refer because she knows her friend will not join).

Note that, now, due to stochastic referral conversion, customers cannot always perfectly infer the size of each priority class by only observing the total queue length unless $p = 1$. Specifically, if the total queue length is greater than n, then all we know is that at least n customers are in the regular class, but we cannot tell the exact number because there could be customers who refer in vain and get stuck in the regular class.

Under threshold-strategy n, let $W(i; n)$ be the expected waiting time of a customer seeing total queue length $i - 1$ upon arrival who does not refer (or refers in vain).

Proposition 3.8. *Customers follow a threshold referral strategy in equilibrium if* $W(i; n) - W(i-1; n) \geq 1/\mu$ *for* $i = 2, \ldots, N-1$.

Note that this if-condition seems intuitive since it says a customer waiting one position behind another customer should expect at least $1/\mu$ more time in the queue due to one more

service completion required. We show a numerical example in Table 3.2 does NOT hold.

Table 3.2: A counter example in which $W(i;n) - W(i-1;n) \geq 1/\mu$ does not hold.

i	$W(i;n)$	$W(i,n) - W(i-1;n)$
1	2.040	
2	5.253	3.213
3	9.305	4.0516
4	11.386	2.0810
5	12.478	1.092
6	13.073	**0.595**
7	13.418	**0.346**
8	13.638	**0.220**
9	13.795	**0.157**
10	14.795	1

Note. $\mu = 1, \Lambda = 1.1, p = 0.9, N = 10,\ n = 3,$

In the example shown in Table 3.2, $1/\mu = 1$; yet for $i = 6, 7, 8, 9$, $W(i,n) - W(i-1;n) < 1/\mu$. The reason why this occurs is that customers seeing longer queues are less likely to be overtaken by future customers because the finite buffer limits the batch size from referrals even further when the queue is long.

Furthermore, we show in Appendix C.3 an instance in which any pure threshold referral strategy does not constitute an equilibrium. The model primitives of that example are $\mu = 1, \Lambda = 1.1, p = 1, N = 10, c = 1, c_r = 10.7$.

3.8 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.

The present chapter 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 credit-

ing 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.

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 fraud is 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.

APPENDIX A

SUPPLEMENT TO CHAPTER 1

A.1 Optimal Direct Mechanism

In this section we study direct mechanisms that maximize the service provider and the intermediary's long-run average revenue, respectively. By the revelation principle, we can restrict our attention without loss of generality to incentive compatible direct mechanism in which each arriving customer truthfully reports her type c (waiting cost) and takes the recommended courses of actions. Thus the direct mechanisms are used to establish the optimality of auctions for the service provider in Proposition 1.2 and for the intermediary in Theorem 1.3, respectively. The proposed auctions implement the outcomes of the optimal direct mechanisms. The proofs of the results in this section are in Appendix A.5.

The mechanism is characterized by three outcome functions: $W : \Xi \to \mathbb{R}_+$, $P : \Xi \mapsto \mathbb{R}$ and $q : \Xi \mapsto \{0, 1\}$, where $W(c)$ is the expected waiting time of a customer who reports c; $P(c)$ is her expected *net* payment to the intermediary; and $q(c) = 1$ if the customer joins, and $q(c) = 0$ otherwise. The direct mechanism $\langle W(c), P(c), q(c)\rangle$ proceeds as follows:

1. The (feasible) mechanism $\langle W(c), P(c), q(c)\rangle$ is publicly announced.

2. Upon each new arrival, c is drawn from F and it is privately observed by the arriving customer, who then reports c' to the intermediary.

3. The customer joins if $q(c') = 1$, and balks otherwise. If she joins, her expected waiting time and expected net payment are $W(c')$ and $P(c')$, respectively.

The expected utility of type c customer if she truthfully reports her type is $U(c) = q(c)[V - cW(c) - p(c)]$. The intermediary's optimal mechanism solves the following problem:

Problem A.1 (Intermediary's Problem)**.**

$$\max_{W(\cdot),P(\cdot),q(\cdot)} \quad \Lambda \int_{\underline{c}}^{\overline{c}} q(c)P(c)dF(c) \tag{A.1a}$$

$$\text{s.t.} \qquad U(c) \geq 0, \qquad \forall c \in \Xi, \tag{A.1b}$$

$$U(c) \geq q(c)\left[V - \frac{c}{\mu - \Lambda \int_{\underline{c}}^{\overline{c}} q(c)dF(c)}\right], \qquad \forall c \in \Xi, \tag{A.1c}$$

$$U(c) \geq q(c')[V - cW(c') - P(c')], \qquad \forall c, c' \in \Xi, \tag{A.1d}$$

$$\frac{\Lambda}{\mu}\int_{c\in\mathcal{C}} q(c)W(c)dF(c) \geq \frac{\frac{\Lambda}{\mu}\int_{c\in\mathcal{C}} q(c)dF(c)}{\mu - \Lambda \int_{c\in\mathcal{C}} q(c)dF(c)}, \qquad \forall \mathcal{C} \subseteq \Xi, \tag{A.1e}$$

$$\frac{\Lambda}{\mu}\int_{\underline{c}}^{\overline{c}} q(c)W(c)dF(c) = \frac{\frac{\Lambda}{\mu}\int_{\underline{c}}^{\overline{c}} q(c)dF(c)}{\mu - \Lambda \int_{\underline{c}}^{\overline{c}} q(c)dF(c)}. \tag{A.1f}$$

Here (A.1a) is the objective function of the intermediary. Constraints (A.1b) are individual rationality (IR) constraints, which ensure that customer c has nonnegative expected utility under if she chooses $(W(c), P(c), q(c))$. Constraints (A.1c) are the property right constraints, which ensure that if a customer c joins the system, then her expected utility under truthful reporting is at least as high as what she would obtain in a FIFO system, i.e., trading is voluntary. Note that if customers do not join the system, this system is irrelevant since $q(c) = 0$. Constraints (A.1d) are the incentive compatibility (IC) constraints, which ensure that customer c reports her type truthfully. Constraints (A.1e) specify the operationally achievable region. The RHS of (A.1e) is the long-run average work in the system under a work conserving policy where all joining customers in the set $\mathcal{C}$ receive strict preemptive priority over all other customers. It is equal to the average work in an $M/M/1$ FIFO system with service rate μ and arrival rate $\Lambda \int_{c\in\mathcal{C}} q(c)dF(c)$. Equation (A.1f) enforces work conservation. Note that the effective arrival rate is $\lambda = \Lambda \int_{\underline{c}}^{\overline{c}} q(c)dF(c)$. We say $W(\cdot) \in OA(\lambda)$ if it is operationally achievable (OA) and work conserving.

As for the service provider's mechanism design problem, constraints (A.1c) become

$$U(c) \geq q(c)\left[V - \frac{c}{\mu - \Lambda \int_{\underline{c}}^{\overline{c}} q(c)dF(c)} - P(c_m)\right], \qquad \forall c \in \Xi,$$

where $W(c_m) = \frac{1}{\mu - \Lambda \int_{\underline{c}}^{\overline{c}} q(c)dF(c)}$. This implies that any joining customer must be at least as well off as if she waits FIFO and makes an expected payment equal to that of a FIFO customer. Notice that this set of constraints is implied by IC constraints (A.1d). Hence Problem A.1'.

Problem A.1' (Service Provider's Problem)**.** The same as (A.1a)-(A.1f) except that (A.1c) is removed.

We first simplify the problem by exploiting the properties of the IC constraints and IR constraints.

Lemma A.1. *The IC and IR constraints hold if and only if*

(i) There exists a threshold $\tilde{c} \in (\underline{c}, \overline{c}]$ *such that only customers with* $c \leq \tilde{c}$ *join, i.e,* $\lambda = \Lambda F(\tilde{c})$*;*

(ii) For $c \in [\underline{c}, \tilde{c}]$*,* $W(c)$ *is weakly decreasing in* c*;* $P(c)$ *is weakly increasing in* c*;* $U(c)$ *is convex decreasing in* c*;*

(iii)

$$P(c_i) - P(c_j) = c_j W(c_j) - c_i W(c_i) - \int_{c_i}^{c_j} W(c)dc, \qquad \forall c_i, c_j \in [\underline{c}, \tilde{c}] \tag{A.2}$$

$$U(c) = U(\tilde{c}) + \int_{c}^{\tilde{c}} W(s)ds, \qquad \forall c \in [\underline{c}, \tilde{c}], \tag{A.3}$$

where $U(\tilde{c}) \geq 0$*, and if* $\tilde{c} < \overline{c}$*,* $U(\tilde{c}) = 0$*.*

Denote $P(\tilde{c})$ by $\tilde{p}$. Lemma A.1 allows us to simplify the service provider's objective

function:

$$\begin{aligned}\Lambda \int_{\underline{c}}^{\overline{c}} q(c)P(c)dF(c) &= \Lambda \int_{\underline{c}}^{\tilde{c}} P(c)dF(c) \\ &= \Lambda \int_{\underline{c}}^{\tilde{c}} \left[\tilde{p} + \tilde{c}W(\tilde{c}) - cW(c) - \int_{c}^{\tilde{c}} W(s)ds\right] dF(c) \\ &= \tilde{p}\Lambda F(\tilde{c}) - \Lambda \int_{\underline{c}}^{\tilde{c}} (W(c) - W(\tilde{c})) \left[c + \frac{F(c)}{f(c)}\right] dF(c),\end{aligned}$$

where the final step is by interchange of integrals. We shall see a similar technique in the manipulation of the intermediary's objective function.

By (A.3), $U(c)$ is continuous and differentiable almost everywhere and $U'(c) = -W(c)$ for $c \in [\underline{c}, \tilde{c}]$ except on a set of measure zero. In the intermediary's problem, we use the "worst-off" types to refer to joining customers that benefit the least in the mechanism relative to being served FIFO. Formally, the "worst-off" types constitute the following set

$$\Theta = \arg \min_{c \in [\underline{c}, \tilde{c}]} \left\{U(c) - \left(V - c\overline{W}(\tilde{c})\right)\right\}.$$

where $\overline{W}(\tilde{c}) = \frac{1}{\mu - \Lambda F(\tilde{c})}$. Since $U(c)$ is convex, Θ is an interval (possibly a singleton) of c at which the subdifferential of $U(c) - \left(V - c\overline{W}(\tilde{c})\right)$ contains 0, i.e.,

$$\Theta = \left\{c | W(c_1) > \overline{W}(\tilde{c}), \forall c_1 < c; W(c_2) < \overline{W}(\tilde{c}), \forall c_2 > c\right\}.$$

Note that due to work conservation, the worst-off types must lie in the interior of $[\underline{c}, \tilde{c}]$, and therefore, Θ is precise. Denote $\Theta = [c_r, c_p]$, where $c_r \leq c_p$. Thus the property right constraints (A.1c) are satisfied as long as they are satisfied for $c \in \Theta$. If Θ is a nonempty interval, i.e., $c_r < c_p$, then $W(c) = \overline{W}(\tilde{c})$ for $c \in [c_r, c_p]$. It is immediate from (A.2) that $P(c) \equiv p_m$ is constant over $[c_r, c_p]$. Thus Constraints (A.1c) simplify to $p_m \leq 0$. If $c_r = c_p = c_m$ and $W(c)$ is differentiable at c_m, then the same result holds. If $c_r = c_p = c_m$ yet $W(c)$ is non-differentiable at c_m, then , we can create an artificial $W(c_m) = \overline{W}(\tilde{c})$ and a

corresponding artificial $P(c_m) = p_m$ such that by (A.2),

$$P(c_m^-) = P(c_m) + c_m \left[W(c_m^-) - \overline{W}(\tilde{c})\right], \quad \text{and} \quad P(c_m^+) = P(c_m) + c_m \left[W(c_m^+) - \overline{W}(\tilde{c})\right].$$

Thus once again the property right constraints (A.1c) simplify to $p_m \leq 0$. This is summarized in the following lemma:

Lemma A.2. *A mechanism that satisfies IC, IR and property right constraints must have*

$$P(c) = \begin{cases} p_m + c_r \overline{W}(\tilde{c}) - cW(c) - \int_c^{c_r} W(s)ds, & c \in [\underline{c}, c_r), \\ p_m + c_p \overline{W}(\tilde{c}) - cW(c) + \int_{c_p}^{c} W(s)ds, & c \in (c_p, \overline{c}]. \end{cases} \tag{A.4}$$

where $p_m \leq 0$. *and* $P(c) = p_m$ *for* $c \in [c_r, c_p]$ *if* $c_r < c_p$.

The intermediary's objective function simplifies to

$$\begin{aligned}
&\Lambda \int_{\underline{c}}^{\overline{c}} q(c)P(c)dF(c) \\
\overset{(1)}{=}&\Lambda \int_{\underline{c}}^{\tilde{c}} P(c)dF(c) \\
\overset{(2)}{=}&\Lambda \int_{\underline{c}}^{c_r} \left[p_m + c_r \overline{W}(\tilde{c}) - cW(c) - \int_c^{c_r} W(s)ds\right] dF(c) \\
&+ \Lambda \int_{c_p}^{\tilde{c}} \left[p_m + c_p \overline{W}(\tilde{c}) - cW(c) + \int_{c_p}^{c} W(s)ds\right] dF(c) + p_m \Lambda (F(c_p) - F(c_r)) \\
=&p_m \Lambda F(\tilde{c}) + \Lambda \int_{\underline{c}}^{c_r} \left[c_r \overline{W}(\tilde{c}) - cW(c) - \int_c^{c_r} W(s)ds\right] dF(c) \\
&+ \Lambda \int_{c_p}^{\tilde{c}} \left[c_p \overline{W}(\tilde{c}) - cW(c) + \int_{c_p}^{c} W(s)ds\right] dF(c) \\
\overset{(3)}{=}&p_m \Lambda F(\tilde{c}) - \Lambda \int_{\underline{c}}^{c_r} (W(c) - \overline{W}(\tilde{c})) \left[c + \frac{F(c)}{f(c)}\right] dF(c) + \Lambda \int_{c_p}^{\tilde{c}} (\overline{W}(\tilde{c}) - W(c)) \left[c - \frac{F(\tilde{c}) - F(c)}{f(c)}\right] dF(c),
\end{aligned}$$

where (1) is by Lemma A.1; (2) is by Lemma A.2; and (3) is by the following observation:

$$c_r \overline{W}(\tilde{c}) F(c_r) = \int_{\underline{c}}^{c_r} \overline{W}(\tilde{c}) \left[c + \frac{F(c)}{f(c)}\right] dF(c),$$

$$c_p\overline{W}(\tilde{c})[F(\tilde{c})-F(c_p)]=\int_{c_p}^{\tilde{c}}\overline{W}(\tilde{c})\left[c-\frac{F(\tilde{c})-F(c)}{f(c)}\right]dF(c),$$

$$\int_{\underline{c}}^{c_r}\int_{c}^{c_r}W(s)dsdF(c)=\int_{\underline{c}}^{c_r}W(c)\int_{\underline{c}}^{c}dF(s)dc=\int_{\underline{c}}^{c_r}W(c)\frac{F(c)}{f(c)}dF(c),$$

$$\int_{c_p}^{\tilde{c}}\int_{c_p}^{c}W(s)dsdF(c)=\int_{c_p}^{\tilde{c}}W(c)\int_{c}^{\tilde{c}}dF(s)dF(c)=\int_{c_p}^{\tilde{c}}W(c)\left[\frac{F(\tilde{c})-F(c)}{f(c)}\right]dF(c).$$

By Lemma A.2,

$$U(\tilde{c})=V-cW(c)-P(c)=V-\int_{c_p}^{\tilde{c}}W(s)ds-c_p\overline{W}(\tilde{c})-p_m.$$

Consequently, the service provider and the intermediary's problems are translated to the following:

Problem A.2 (Intermediary's Problem)**.**

$$\max_{W(\cdot),c_r,c_p,\tilde{c}\in\Xi,p_m\leq 0}\quad p_m\Lambda F(\tilde{c})-\Lambda\int_{\underline{c}}^{c_r}(W(c)-\overline{W}(\tilde{c}))f_r(c)dF(c)+\Lambda\int_{c_p}^{\tilde{c}}(\overline{W}(\tilde{c})-W(c))f_p(c;\tilde{c})dF(c) \tag{A.5a}$$

$$\text{s.t.}\quad V-\int_{c_p}^{\tilde{c}}W(c)dc-c_p\overline{W}(\tilde{c})-p_m\geq 0, \tag{A.5b}$$

$$W(\cdot)\ \textit{nonincreasing and}\ W(\cdot)\in OA(\Lambda F(\tilde{c})), \tag{A.5c}$$

$$c_r\leq c_p\leq\tilde{c}. \tag{A.5d}$$

Problem A.2' (Service Provider's Problem)**.**

$$\max_{W(\cdot),\tilde{c}\in\Xi,\tilde{p}\in\mathbb{R}}\quad \tilde{p}\Lambda F(\tilde{c})-\Lambda\int_{\underline{c}}^{\tilde{c}}\left[W(c)-W(\tilde{c})\right]f_r(c)dF(c) \tag{A.6a}$$

$$\text{s.t.}\quad V-W(\tilde{c})-\tilde{p}\geq 0, \tag{A.6b}$$

$$W(\cdot)\ \text{nonincreasing and}\ W(\cdot)\in OA(\Lambda F(\tilde{c})). \tag{A.6c}$$

In our analyses of the intermediary's problem, we first identify the structure of the waiting time function for any given $\tilde{c}$. This allows for further simplification of the problem. Then

we solve for the optimal $\tilde{c}$ using the structural properties derived. The service problem is simpler and can be similarly handled.

Lemma A.3. *For the intermediary's problem, fix $\tilde{c}$ with $\Lambda F(\tilde{c}) < \mu$.*

(i) The optimal expected waiting time is

$$W^*(c;\tilde{c}) = \begin{cases} W^e(c;\tilde{c}) > \overline{W}(\tilde{c}), & \forall\ c \in [\underline{c}, c_r^*(\tilde{c})); \\ \overline{W}(\tilde{c}), & \forall c \in [c_r^*(\tilde{c}), c_p^*(\tilde{c})]. \\ W^e(c;\tilde{c}) < \overline{W}(\tilde{c}), & \forall\ c \in (c_p^*(\tilde{c}), \bar{c}]. \end{cases} \tag{A.7}$$

(ii) $c_r^*(\tilde{c})$ and $c_p^*(\tilde{c})$ satisfy

$$\frac{1}{\mu(1-\rho F(\tilde{c})+\rho F(c_r^*(\tilde{c})))(1-\rho F(\tilde{c})+\rho F(c_p^*(\tilde{c})))} = \overline{W}(\tilde{c}). \tag{A.8}$$

(iii) $f_r(c_r^*(\tilde{c})) \geq f_p(c_p^*(\tilde{c});\tilde{c})$. Particularly, if (A.5b) is slack, then $f_r(c_r^*(\tilde{c})) = f_p(c_p^*(\tilde{c});\tilde{c})$.

(iv) If $c_r^*(\tilde{c}) < c_p^*(\tilde{c})$, then $p_m^*(\tilde{c}) = 0$.

Lemma A.4. *For the service provider's problem, fix $\tilde{c}$ with $\Lambda F(\tilde{c}) < \mu$, the optimal waiting time is $W^e(c;\tilde{c})$.*

In the intermediary's problem, when $\tilde{c}$ is fixed, $p_m^*(\tilde{c})$ could be strictly negative, and Lemma A.3-(iv) implies this happens only if $c_r^*(\tilde{c}) = c_p^*(\tilde{c})$. If $\tilde{c}$ is also optimally determined, Lemma A.5 shows that $p_m^* = 0$ must hold.

Lemma A.5. *In the intermediary's optimal mechanism, the worst-off types' expected net payment is zero , i.e., $p_m^* = 0$. Hence, $U(c) = V - \frac{c}{\mu - \Lambda F(\tilde{c})},\ c \in [c_r, c_p]$.*

From Lemma A.3 and A.5, we can further simplify the intermediary's mechanism design problem to the following:

Problem A.3 (Intermediary's Problem)**.**

$$\max_{c_r,c_p,\tilde{c}\in\Xi:c_r\le c_p\le\tilde{c}} \quad -\Lambda\int_{\underline{c}}^{c_r}\left(W^e(c;\tilde{c})-\overline{W}(\tilde{c})\right)f_r(c)dF(c)+\Lambda\int_{c_p}^{\tilde{c}}\left(\overline{W}(\tilde{c})-W^e(c;\tilde{c})\right)f_p(c;\tilde{c})dF(c) \tag{A.9a}$$

$$\text{s.t.} \quad \frac{1}{\mu(1-\rho F(\tilde{c})+\rho F(c_r))(1-\rho F(\tilde{c})+\rho F(c_p))}=\overline{W}(\tilde{c}), \tag{A.9b}$$

$$V-\int_{c_p}^{\tilde{c}}W^e(c;\tilde{c})dc-c_p\overline{W}(\tilde{c})\ge 0. \tag{A.9c}$$

Problem A.3 is the final version of the intermediary's problem we work with. It appears in Theorem 1.3. Note that Assumption 1.1 guarantees the feasibility of Problem A.3 since it admits a solution $c_r=c_p=\tilde{c}=\underline{c}$.

Problem A.3' (Service Provider's Problem)**.**

$$\max_{\tilde{c}\in\Xi,\tilde{p}\in\mathbb{R}} \quad \tilde{p}\Lambda F(\tilde{c})-\Lambda\int_{\underline{c}}^{\tilde{c}}\left[W^e(c;\tilde{c})-\frac{1}{\mu}\right]f_r(c)dF(c) \tag{A.10a}$$

$$\text{s.t.} \quad V-\frac{\tilde{c}}{\mu}-\tilde{p}\ge 0. \tag{A.10b}$$

Problem A.3' is the final version of the service provider's problem we work with. The baseline auction and the admission fee p^M in Proposition 1.2 implements the outcome of this problem. Again, Assumption 1.1 guarantees the feasibility of Problem A.3' since it admits a solution $\tilde{c}=\underline{c},\ \tilde{p}=0$.

A.2 Proofs of Results in Section 1.4

Proof of Theorem 1.1. The proof consists of five steps. We start with an equilibrium conjecture that there exists a cutoff delay cost $\tilde{c}$ such that only customers with type $c\in[\underline{c},\tilde{c}]$ join the system, and that for joining customers $c\in[\underline{c},\tilde{c}]$, there exists a symmetric, strictly increasing equilibrium bidding strategy $b^{\mathrm{B}}(c;\tilde{c})$. We derive the equilibrium waiting time and bid function based on this conjecture in Step 1. We verify that this constitutes an

equilibrium in Step 2-5. Specifically, part (i) is conjectured in Step 1 and verified in Step 4; the equilibrium bid and waiting time functions in part (ii) are derived in Step 1 and verified in subsequent steps; in particular, Step 2 verifies that the equilibrium bid function is strictly increasing; the first half of part (iii) is shown in Step 3, the second half of part (iii) is shown in Step 4.

Step 1: Equilibrium bid and waiting time functions. Similar to Kleinrock (1967), if the equilibrium bidding strategy is strictly increasing, the expected (efficient) waiting time for $c \in [\underline{c}, \tilde{c}]$ given the arrival rate $\lambda = \Lambda F(\tilde{c})$ is

$$W^e(c;\tilde{c}) = \frac{1}{\mu(1-\rho F(\tilde{c})+\rho F(c))^2}. \tag{A.11}$$

We concisely reproduce its derivation in our context. Given the arrival rate $\lambda = \Lambda F(\tilde{c})$, for customer $c \in [\underline{c}, \tilde{c}]$, her expected waiting time $W^e(c;\tilde{c})$ has three contributions: she expects FIFO waiting time $\frac{1}{\mu-\lambda}$ upon arrival; she expects her waiting time to be shortened by buying from existing customers in $[\underline{c}, c]$, $\int_{\underline{c}}^{c} \frac{1}{\mu}\Lambda W^e(s;\tilde{c})dF(s)$; she expects her waiting time to be lengthened by selling to future customers in $[c, \tilde{c}]$ who later arrive during her stay in the system, $\frac{1}{\mu}W^e(c,\tilde{c})\Lambda[F(\tilde{c})-F(c)]$. Thus,

$$W^e(c;\tilde{c}) = \frac{1}{\mu-\lambda} - \int_{\underline{c}}^{c} \frac{1}{\mu}\Lambda W^e(s;\tilde{c})dF(s) + \frac{1}{\mu}W^e(c,\tilde{c})\Lambda[F(\tilde{c})-F(c)]. \tag{A.12}$$

Letting $c = \underline{c}$ in (A.12) gives the boundary condition $W^e(\underline{c};\tilde{c}) = \frac{1}{\mu[1-\rho F(\tilde{c})]^2}$. Taking the derivative on both sides of (A.12) w.r.t c gives a differential equation. Solving this differential equation gives (A.11).

We first fix $\tilde{c}$ and use the shorthand notation $b(\cdot)$ for the bid function and $W^e(\cdot)$ for the efficient waiting time function. We derive the bidding function for any given $\tilde{c}$ and then pin

down $\tilde{c}$ in Step 3. The expected utility of customer c who bids $\beta \in [b(\underline{c}), b(\tilde{c})]$ is

$$U(c,\beta) = V - cW^e(b^{-1}(\beta)) - \underbrace{\int_{\underline{c}}^{b^{-1}(\beta)} \frac{b(s)}{\mu}\Lambda W^e(s)dF(s)}_{P_p(\beta)} + \underbrace{\frac{\beta}{\mu}W^e(b^{-1}(\beta))\Lambda[F(\tilde{c}) - F(b^{-1}(\beta))]}_{P_r(\beta)}.$$

The best response for customer c solves

$$\begin{aligned}\frac{\partial U}{\partial \beta} = &- \frac{cW^{e\prime}(b^{-1}(\beta))}{b'(b^{-1}(\beta))} - \frac{\rho\beta W^e(b^{-1}(\beta))f(b^{-1}(\beta))}{b'(b^{-1}(\beta))} + \frac{\rho\beta W^e(b^{-1}(\beta))[-f(b^{-1}(\beta))]}{b'(b^{-1}(\beta))} \\ &+ \rho[W^e(b^{-1}(\beta)) + \frac{\beta W^{e\prime}(b^{-1}(\beta))}{b'(b^{-1}(\beta))}][F(\tilde{c}) - F(b^{-1}(\beta))] = 0.\end{aligned}$$

In equilibrium, $\beta = b(c)$. Thus, the symmetric equilibrium $b(c)$ satisfies the following

$$-cW^{e\prime}(c) - 2\rho b(c)W^e(c)f(c) + \rho[W^e(c)b'(c) + b(c)W^{e\prime}(c)][F(\tilde{c}) - F(c)] = 0.$$

After some algebra, this simplifies to the following linear differential equation:

$$[b'(c) - 1]\frac{(F(\tilde{c}) - F(c))^2}{(1 - \rho F(\tilde{c}) + \rho F(c))^2} - [b(c) - c]\frac{2f(c)(F(\tilde{c}) - F(c))}{(1 - \rho F(\tilde{c}) + \rho F(c))^3} = -\frac{(F(\tilde{c}) - F(c))^2}{(1 - \rho F(\tilde{c}) + \rho F(c))^2}. \tag{A.13}$$

The solution to (A.13) is given by

$$[b(c) - c]\frac{(F(\tilde{c}) - F(c))^2}{(1 - \rho F(\tilde{c}) + \rho F(c))^2} = \int_{\tilde{c}}^{c} -\frac{(F(\tilde{c}) - F(s))^2}{(1 - \rho F(\tilde{c}) + \rho F(s))^2}ds + K,$$

where the constant K is chosen such that the right hand side is equal to zero at $c = \tilde{c}$; otherwise, $b(c)$ goes to infinity as c approaches $\tilde{c}$. This implies $K = 0$. Therefore,

$$b(c) = c + \frac{(1 - \rho F(\tilde{c}) + \rho F(c))^2}{(F(\tilde{c}) - F(c))^2}\int_{c}^{\tilde{c}} \frac{(F(\tilde{c}) - F(s))^2}{(1 - \rho F(\tilde{c}) + \rho F(s))^2}ds.$$

By L'Hospital's rule,

$$\lim_{c\to\tilde{c}} \frac{\int_c^{\tilde{c}} \frac{(F(\tilde{c})-F(s))^2}{(1-\rho F(\tilde{c})+\rho F(s))^2} ds}{(F(\tilde{c})-F(c))^2} = \lim_{c\to\tilde{c}} \frac{-\frac{(F(\tilde{c})-F(c))^2}{(1-\rho F(\tilde{c})+\rho F(c))^2}}{-2(F(\tilde{c})-F(c))f(c)} = 0.$$

Thus, $\lim_{c\to\tilde{c}} b(c) = \tilde{c}$.

Step 2: Monotonicity. Now we shall show the monotonicity of the bidding function. Since $b(c) > c$ for $c \in [\underline{c}, \tilde{c})$, from differential equation (A.13) we have

$$b'(c) = [b(c) - c]\frac{2f(c)}{(1-\rho F(\tilde{c})+\rho F(c))(F(\tilde{c})-F(c))} > 0.$$

Therefore, the assumption of a strictly increasing bidding strategy is verified. The equilibrium achieves efficient scheduling.

Step 3: IR in trading. It remains to verify that all joining customers are at least as well off by submitting a bid and participating in trading as if she just waits FIFO. That is $\vartheta(c) \equiv U(c, b(c)) - [V - \frac{c}{\mu-\lambda}] \geq 0$ for all $c \in [\underline{c}, \tilde{c}]$, where $\lambda = \Lambda F(\tilde{c})$.

$$U(c, b(c)) = V - cW^e(c;\tilde{c}) - \rho\int_{\underline{c}}^{c} b(s)W^e(s;\tilde{c})dF(s) + \rho W^e(c;\tilde{c})b(c)(F(\tilde{c}) - F(c))$$

By the envelope theorem,

$$\frac{dU(c, b(c))}{dc} = -W^e(c;\tilde{c}) < 0.$$

Further differentiating w.r.t. c yields

$$\frac{d^2U(c, b(c))}{dc^2} = -\frac{\partial W^e(c;\tilde{c})}{\partial c} > 0.$$

Hence, $U(c, b(c))$ is convex decreasing in c. By applying the first order condition, the minimum of $\vartheta(c)$ is attained at

$$c_m : \frac{1}{\mu(1-\rho F(\tilde{c})+\rho F(c_m))^2} = \frac{1}{\mu-\lambda}$$

Thus, to show IR in trading it suffices to show $\vartheta(c_m) > 0$, i.e.,

$$-\rho \int_{\underline{c}}^{c_m} b(s) W^e(s;\tilde{c}) dF(s) + \rho W^e(c_m;\tilde{c}) b(c_m)(F(\tilde{c}) - F(c_m)) \geq 0.$$

Since $b(c)$ is strictly increasing, it follows that

$$-\rho \int_{\underline{c}}^{c_m} b(s) W^e(s;\tilde{c}) dF(s) > -\rho b(c_m) \int_{\underline{c}}^{c_m} W^e(s;\tilde{c}) dF(s).$$

Since

$$\rho \int_{\underline{c}}^{c_m} W^e(s;\tilde{c}) dF(s) = \frac{1}{\mu(1-\rho F(\tilde{c}))} - \frac{1}{\mu(1-\rho F(\tilde{c})+\rho F(c_m))} = \frac{1}{\mu}\left[\frac{1}{1-\rho F(\tilde{c})} - \frac{1}{\sqrt{1-\rho F(\tilde{c})}}\right],$$

and

$$\rho W^e(c_m;\tilde{c})(F(\tilde{c}) - F(c_m)) = \frac{\rho(F(\tilde{c}) - F(c_m))}{\mu(1-\rho F(\tilde{c}))} = \frac{1-\sqrt{1-\rho F(\tilde{c})}}{\mu(1-\rho F(\tilde{c}))},$$

thus,

$$-\rho b(c_m) \int_{\underline{c}}^{c_m} W^e(s;\tilde{c}) dF(s) + \rho W^e(c_m;\tilde{c}) b(c_m)(F(\tilde{c}) - F(c_m)) = 0.$$

Therefore, $\vartheta(c_m) > 0$.

Step 4: IR in joining. Since $U(c, b(c))$ is decreasing, if $\tilde{c} < \overline{c}$, it must be that $\tilde{c}$ solves $U(\tilde{c}, b(\tilde{c};\tilde{c})) = 0$, i.e., the marginal customer must be indifferent about whether to join the queue. When $\tilde{c} < \overline{c}$, if a customer with $c > \tilde{c}$ bids $b(\hat{c})$ for any $\hat{c} \in [\underline{c}, \tilde{c}]$, she receives a strictly lower utility than customer $\tilde{c}$ would if she bids $b(\hat{c})$. Since the utility of customer $\tilde{c}$ is zero in the best response, a customer with $c > \tilde{c}$ prefers balking to submitting $b(\hat{c})$. Likewise, since customer $\tilde{c}$ prefers trading to joining the queue without trading, customer c prefers balking to joining without trading. Similarly, customer c who bids a real-numbered price outside $[b(\underline{c}), b(\tilde{c})]$ has a lower utility than if she balks. Therefore, a customer with type $c > \tilde{c}$ prefers balking.

Now we show the uniqueness of $\tilde{c}$. First note that $U(\underline{c}, b(\underline{c};\underline{c})) = V - \frac{\underline{c}}{\mu} > 0$ by Assumption

1.1. Thus it suffices to show $U(\tilde{c}, b(\tilde{c};\tilde{c}))$ is strictly decreasing in $\tilde{c}$.

$$U(\tilde{c}, b(\tilde{c};\tilde{c})) = V - \frac{\tilde{c}}{\mu} - \rho \int_{\underline{c}}^{\tilde{c}} b(s;\tilde{c}) W^e(s;\tilde{c}) dF(s).$$

Hence, it suffices to show $P^B(\tilde{c}) = \rho \int_{\underline{c}}^{\tilde{c}} b(s;\tilde{c}) W^e(s;\tilde{c}) dF(s)$ is increasing in $\tilde{c}$. We first make some manipulation on $b(c;\tilde{c})$ to eventually obtain a simplified expression of $P^B(\tilde{c})$.

$$\begin{aligned}
b(c;\tilde{c}) &= c + \frac{\int_c^{\tilde{c}} (F(\tilde{c}) - F(s))^2 W^e(s;\tilde{c}) ds}{(F(\tilde{c}) - F(c))^2 W^e(c;\tilde{c})} \\
&= \frac{c(F(\tilde{c}) - F(c))^2 W^e(c;\tilde{c})}{(F(\tilde{c}) - F(c))^2 W^e(c;\tilde{c})} + \frac{\int_c^{\tilde{c}} (F(\tilde{c}) - F(s))^2 W^e(s;\tilde{c}) ds}{(F(\tilde{c}) - F(c))^2 W^e(c;\tilde{c})} \\
&= \{ \int_c^{\tilde{c}} s(F(\tilde{c}) - F(s))^2 \left[-\frac{\partial W^e(s;\tilde{c})}{\partial s} \right] ds + \int_c^{\tilde{c}} 2s(F(\tilde{c}) - F(s)) f(s) W^e(s;\tilde{c}) ds \\
&\quad - \int_c^{\tilde{c}} (F(\tilde{c}) - F(s))^2 W^e(s;\tilde{c}) ds \} \cdot \left[(F(\tilde{c}) - F(c))^2 W^e(c;\tilde{c}) \right]^{-1} \\
&\quad + \frac{\int_c^{\tilde{c}} (F(\tilde{c}) - F(s))^2 W^e(s;\tilde{c}) ds}{(F(\tilde{c}) - F(c))^2 W^e(c;\tilde{c})} \quad \text{(integration by parts)} \\
&= \frac{\int_c^{\tilde{c}} \left[s(F(\tilde{c}) - F(s))^2 \left[-\frac{\partial W^e(s;\tilde{c})}{\partial s} \right] + 2s(F(\tilde{c}) - F(s)) f(s) W^e(s;\tilde{c}) \right] ds}{(F(\tilde{c}) - F(c))^2 W^e(c;\tilde{c})} \\
&= \frac{\int_c^{\tilde{c}} \left[\rho s(F(\tilde{c}) - F(s))^2 + s(F(\tilde{c}) - F(s)) \left(1 - \rho F(\tilde{c}) + \rho F(s)\right) \right] \left[-\frac{\partial W^e(s;\tilde{c})}{\partial s} \right] ds}{\rho (F(\tilde{c}) - F(c))^2 W^e(c;\tilde{c})} \\
&= \frac{\int_c^{\tilde{c}} s(F(\tilde{c}) - F(s)) \left[-\frac{\partial W^e(s;\tilde{c})}{\partial s} \right] ds}{\rho (F(\tilde{c}) - F(c))^2 W^e(c;\tilde{c})}.
\end{aligned}$$

Substituting $b(c;\tilde{c})$ into $P^B(\tilde{c})$ gives

$$\begin{aligned}
P^B(\tilde{c}) &= \rho\int_{\underline{c}}^{\tilde{c}} b(s;\tilde{c})W^e(s;\tilde{c})dF(s) \\
&= \rho\int_{\underline{c}}^{\tilde{c}} \left[\frac{-\int_s^{\tilde{c}} x(F(\tilde{c})-F(x))\frac{\partial W^e(x;\tilde{c})}{\partial x}dx}{\rho(F(\tilde{c})-F(s))^2 W^e(s;\tilde{c})}\right] W^e(s;\tilde{c})dF(s) \\
&= \int_{\underline{c}}^{\tilde{c}} \frac{-\int_s^{\tilde{c}} x(F(\tilde{c})-F(x))\frac{\partial W^e(x;\tilde{c})}{\partial x}dx}{(F(\tilde{c})-F(s))^2}dF(s) \\
&= -\int_{\underline{c}}^{\tilde{c}} x(F(\tilde{c})-F(x))\frac{\partial W^e(x;\tilde{c})}{\partial x}\int_{\underline{c}}^{x}\frac{1}{(F(\tilde{c})-F(s))^2}dF(s)dx \quad \text{(interchange of integrals)} \\
&= -\int_{\underline{c}}^{\tilde{c}} x(F(\tilde{c})-F(x))\frac{\partial W^e(x;\tilde{c})}{\partial x}\left[\frac{1}{(F(\tilde{c})-F(x))}-\frac{1}{F(\tilde{c})}\right]dx \\
&= \int_{\underline{c}}^{\tilde{c}} -\frac{\partial W^e(x;\tilde{c})}{\partial x}\left[x-x+\frac{xF(x)}{F(\tilde{c})}\right]dx \\
&= \frac{\int_{\underline{c}}^{\tilde{c}} cF(c)\left[-\frac{\partial W^e(c;\tilde{c})}{\partial c}\right]dc}{F(\tilde{c})} \qquad (A.14) \\
&= \frac{\int_{\underline{c}}^{\tilde{c}}\left[W^e(c;\tilde{c})-\frac{1}{\mu}\right]f_r(c)dF(c)}{F(\tilde{c})} \quad \text{(integration by parts).} \qquad (A.15)
\end{aligned}$$

Taking the derivative of $P^B(\tilde{c})$ gives

$$\frac{dP^B(\tilde{c})}{d\tilde{c}} = \frac{\int_{\underline{c}}^{\tilde{c}}\left[\frac{\partial W^e(c;\tilde{c})}{\partial\tilde{c}}\right]f_r(c)dF(c)F(\tilde{c}) - f(\tilde{c})\int_{\underline{c}}^{\tilde{c}}\left[W^e(c;\tilde{c})-\frac{1}{\mu}\right]f_r(c)dF(c)}{F^2(\tilde{c})}.$$

To show $\frac{dP^B(\tilde{c})}{d\tilde{c}} > 0$, it suffices to show the numerator is greater than zero.

$$\begin{aligned}
&\int_{\underline{c}}^{\tilde{c}}\left[\frac{\partial W^e(c;\tilde{c})}{\partial\tilde{c}}\right]f_r(c)dF(c)F(\tilde{c}) - f(\tilde{c})\int_{\underline{c}}^{\tilde{c}}\left[W^e(c;\tilde{c})-\frac{1}{\mu}\right]f_r(c)dF(c) \\
= &\int_{\underline{c}}^{\tilde{c}}\left[\frac{2\rho F(\tilde{c})}{\mu(1-\rho F(\tilde{c})+\rho F(c))^3} - \frac{1-\rho F(\tilde{c})+\rho F(c)}{\mu(1-\rho F(\tilde{c})+\rho F(c))^3} + \frac{1}{\mu}\right]f(\tilde{c})f_r(c)dF(c) \\
= &\int_{\underline{c}}^{\tilde{c}}\left[\frac{2F(c)+2\rho(F(\tilde{c})-F(c))^2+\rho(F(\tilde{c})-F(c))^2(1-\rho F(\tilde{c})+\rho F(c))}{\mu(1-\rho F(\tilde{c})+\rho F(c))^3}\right]\rho f(\tilde{c})f_r(c)dF(c) \\
> &\ 0.
\end{aligned}$$

Step 5: IC. It remains to verify that $b(\cdot)$ is indeed an equilibrium strategy from which

no one deviates. This is done by checking the condition that

$$U(c, b(c)) \geq U(c, \beta) \quad \forall c \in [\underline{c}, \tilde{c}], \beta \neq b(c) \tag{A.16}$$

fixing $b(\cdot)$ for the rest of customers. If $\beta < b(\underline{c})$ then bidding $b(\underline{c})$ will not change the expected waiting time but increase the expected payment received as the trading price is determined by the seller's bid, so the customer's expected utility will be increased. Similarly, if $\beta > b(\tilde{c})$, then the customer will be as well off as bidding $b(\tilde{c})$, so bidding $b(\tilde{c})$ is supported in equilibrium. Therefore, to show (A.16), it suffices to show that customer c does not have an incentive to deviate to a bid $b \in [b(\underline{c}), b(\tilde{c})]$. Since $b(\cdot)$ is strictly increasing, we can find a unique c' such that $\beta = b(c')$ for $\beta \in [b(\underline{c}), b(\tilde{c})]$. Notice that

$$\begin{aligned} U(c, \beta) = U(c, b(c')) &= U(c', b(c')) - cW^e(c') + c'W^e(c') \\ &= U(c, b(c)) - \int_c^{c'} W^e(s)ds - cW^e(c') + c'W^e(c'). \end{aligned}$$

Thus, it suffices to show $-\int_c^{c'} W^e(s)ds - cW^e(c') + c'W^e(c') \leq 0$. Since $W^e(\cdot)$ is a decreasing function,

$$-\int_c^{c'} W^e(s)ds - cW^e(c') + c'W^e(c') = \int_c^{c'} [W^e(c') - W^e(s)]ds \leq 0. \qquad \square$$

A.3 Proofs of Results in Section 1.5

Proof of Proposition 1.1. **Part 1** (overjoining): $\lambda^{SW} = \Lambda F(\tilde{c}^{SW})$, where $\tilde{c}^{SW}$ (if less than $\overline{c}$) is determined by the first order condition of (1.3):

$$\Lambda f(\tilde{c}) \left[V - \frac{\tilde{c}}{\mu} \right] - \Lambda \int_{\underline{c}}^{\tilde{c}} c \frac{\partial W^e(c; \tilde{c})}{\partial \tilde{c}} dF(c) = 0.$$

Factoring out $\Lambda f(\tilde{c})$, this is equivalent to

$$V - \frac{\tilde{c}}{\mu} - \int_{\underline{c}}^{\tilde{c}} c \left[-\frac{\partial W^e(c;\tilde{c})}{\partial c} \right] dc = 0. \tag{A.17}$$

Note that $\int_{\underline{c}}^{\tilde{c}} c \left[-\frac{\partial W^e(c;\tilde{c})}{\partial c} \right] dc$ increases in $\tilde{c}$. Thus, the LHS of (A.17) is decreasing in $\tilde{c}$ and $\tilde{c}^{SW} = \overline{c}$ iff the LHS of (A.17) is positive for $\tilde{c} = \overline{c}$.

$\lambda^B = \Lambda F(\tilde{c}^B)$ where $\tilde{c}^B$ solves

$$V - \frac{\tilde{c}}{\mu} - P^B(\tilde{c}) = 0, \tag{A.18}$$

where $P^B(\tilde{c})$ is defined in (A.14), and also increases in $\tilde{c}$ as shown in the proof of Theorem 1.1. Thus, the LHS of (A.18) is decreasing in $\tilde{c}$ and $\tilde{c}^B = \overline{c}$ iff the the LHS of (A.18) is positive for $\tilde{c} = \overline{c}$. Since

$$\int_{\underline{c}}^{\tilde{c}} c \left[-\frac{\partial W^e(c;\tilde{c})}{\partial c} \right] dc - P^B(\tilde{c}) = \int_{\underline{c}}^{\tilde{c}} c \left[1 - \frac{F(c)}{F(\tilde{c})} \right] \left[-\frac{\partial W^e(c;\tilde{c})}{\partial c} \right] dc > 0, \tag{A.19}$$

therefore, comparing (A.17) and (A.18) gives $\tilde{c}^B \geq \tilde{c}^{SW}$ with equality iff $\tilde{c}^B = \tilde{c}^{SW} = \overline{c}$.

Part 2 (socially optimal admission fee): We first show that charging an admission fee will only change the joining function $J(c)$ (different $\tilde{c}$), but not the bid function $b^B(c;\tilde{c})$. Then we find the expression p^{SW}.

Recall that in the baseline auction without the admission fee, all joining customers trade since $U(c, b(c)) \geq V - \frac{c}{\mu - \lambda}$ for $c \in [\underline{c}, \tilde{c}]$. With an admission fee p, this still holds since $U(c, b(c)) - p \geq V - \frac{c}{\mu - \lambda} - p$ for $c \in [\underline{c}, \tilde{c}]$, where $U(c, b(c)) - p$ is the expected utility of trading (which implies joining the system and paying p). Therefore, since the bidding function is not affected, the expected waiting time is still $W^e(c;\tilde{c})$. Thus, the expression for p^{SW} follows from (A.19). □

Proof of Lemma 1.1. The sufficient part is immediate. We show the necessary part. Suppose that $\frac{df_p(c;\overline{c})}{dc} > 0$ for all $c \in [\underline{c}, \overline{c}]$, i.e., $1 - \frac{-f^2(c) - (1-F(c))f'(c)}{f^2(c)} > 0, \ \forall c \in [\underline{c}, \overline{c}]$.

Equivalently,

$$2 + \frac{(1 - F(c))f'(c)}{f^2(c)} > 0, \quad \forall c \in [\underline{c}, \overline{c}].$$

For any given c, if $f'(c) > 0$, then for any $x \in [F(c), 1]$,

$$v(x) \triangleq 2 + \frac{(x - F(c))f'(c)}{f^2(c)} > 0.$$

If $f'(c) < 0$, then for $x \in [F(c), 1]$,

$$v(x) \geq 2 + \frac{(1 - F(c))f'(c)}{f^2(c)} > 0.$$

Therefore, $\frac{df_p(c;\tilde{c})}{dc} = v(F(\tilde{c})) > 0$. □

Proof of Proposition 1.2. The service provider's optimal direct mechanism solves Problem A.3' in Appendix A.1. We shall show the baseline auction plus an admission fee p^M implement the servicer provider's optimal direct mechanism and thus raises the maximum revenue. First, we know that expected waiting time under the baseline auction is $W^e(c; \tilde{c})$, the same as that of the optimal direct mechanism (see Lemma A.3') for a given $\tilde{c}$. Moreover, as we argue in Proposition 1.1, the admission fee does not change the equilibrium bid function. Second, no customers would deviate from the equilibrium bid function $b^B(c; \tilde{c})$, which corresponds to IC in the direct mechanism. Finally, if p^M is set such that the arrival rate of the auction matches that in the optimal direct mechanism in Problem A.3', then by the revenue equivalence theorem, the proposed auction maximizes the service provider's revenue.

First note that constraint (A.10b) must be binding. Suppose it is slack, then one can always increase $\tilde{p}$ by ϵ to strictly increase the objective function without violating the constraint. Hence, substituting $\tilde{p} = V - \frac{\tilde{c}}{\mu}$ in the objective function gives

$$\Lambda F(\tilde{c}) \left(V - \frac{\tilde{c}}{\mu} \right) - \Lambda \int_{\underline{c}}^{\tilde{c}} \left[W^e(c; \tilde{c}) - \frac{1}{\mu} \right] f_r(c) dF(c).$$

Case 1: If $\tilde{c}^M < \overline{c}$, it solves the first order condition

$$\Lambda f(\tilde{c}) \left[V - \frac{\tilde{c}}{\mu} \right] + \Lambda F(\tilde{c}) \left(-\frac{1}{\mu} \right) - \Lambda \int_{\underline{c}}^{\tilde{c}} \left[\frac{\partial W^e(c;\tilde{c})}{\partial \tilde{c}} \right] f_r(c) dF(c) = 0.$$

Factoring out $\Lambda f(\tilde{c})$ gives

$$V - \frac{1}{\mu} \left[\tilde{c} + \frac{F(\tilde{c})}{f(\tilde{c})} \right] - \frac{2\rho}{\mu} \int_{\underline{c}}^{\tilde{c}} \frac{cf(c) + F(c)}{[1 - \rho F(\tilde{c}) + \rho F(c)]^3} dc = 0. \tag{A.20}$$

By Assumption 1.2, $\tilde{c} + \frac{F(\tilde{c})}{f(\tilde{c})}$ increases in $\tilde{c}$. It is also easy to see $\frac{2\rho}{\mu} \int_{\underline{c}}^{\tilde{c}} \frac{cf(c)+F(c)}{[1-\rho F(\tilde{c})+\rho F(c)]^3} dc$ increases in $\tilde{c}$. Thus the LHS of (A.20) decreases in $\tilde{c}$. Recall that the LHS of (A.17) also decreases in $\tilde{c}$. Comparing (A.17) and (A.20), since

$$\int_{\underline{c}}^{\tilde{c}} c \left[-\frac{\partial W^e(c;\tilde{c})}{\partial c} \right] dc = \frac{2\rho}{\mu} \int_{\underline{c}}^{\tilde{c}} \frac{cf(c)}{[1 - \rho F(\tilde{c}) + \rho F(c)]^3} dc < \frac{2\rho}{\mu} \int_{\underline{c}}^{\tilde{c}} \frac{cf(c) + F(c)}{[1 - \rho F(\tilde{c}) + \rho F(c)]^3} dc,$$

it follows that $\tilde{c}^M < \tilde{c}^{SW}$. Since the maximum revenue should be no less than the revenue raised in the socially optimal mechanism, i.e., $p^M \Lambda F(\tilde{c}^M) \geq p^{SW} \Lambda F(c^{SW})$. it follow that $p^M > p^{SW}$.

Case 2: If $\tilde{c}^M = \overline{c}$, which implies the LHS of (A.20) is positive when $\tilde{c} = \overline{c}$; hence, the LHS of (A.17) is also positive.

$$V - \frac{\overline{c}}{\mu} - \int_{\underline{c}}^{\overline{c}} c \left[-\frac{\partial W^e(c;\overline{c})}{\partial c} \right] dc > 0. \tag{A.21}$$

Therefore, $\tilde{c}^{SW} = \tilde{c}^M = \overline{c}$. Recall that (A.10b) must be binding, and thus

$$V - \frac{\overline{c}}{\mu} - P^B(\overline{c}) - p^M = 0. \tag{A.22}$$

From (A.19),

$$P^{SW} = \int_{\underline{c}}^{\overline{c}} c \left[-\frac{\partial W^e(c;\overline{c})}{\partial c} \right] dc - P^B(\overline{c}). \tag{A.23}$$

Combining (A.21), (A.22) and (A.23) yields $p^M > p^{SW}$. □

A.4 Proofs of Results in Section 1.6

Proof of Theorem 1.2. The proof consists of four steps. We start with an equilibrium conjecture that there exists a cutoff delay cost $\tilde{c}$ such that only customers with type $c \in [\underline{c}, \tilde{c}]$ join the system (part (i)), and that for joining customers $c \in [\underline{c}, \tilde{c}]$, there exists a symmetric, symmetric bidding strategy $b^A(c; c_r, c_p, \tilde{c})$ that is strictly increasing in $[\underline{c}, c_r)$ and $(c_p, \tilde{c}]$ respectively with $b^A(c_r^-; c_r, c_p, \tilde{c}) \leq \underline{R}$. Customers in $[c_r, c_p]$ bid $\overline{R}$ with $b^A(c_p^+, c_r, c_p, \tilde{c}) \geq \overline{R}$. We derive the equilibrium bid function (part (ii)) and waiting time (part (iii)) based on this conjecture in Step 1. We verify that this constitutes an equilibrium in Step 2-4. Specifically, (1.6a)-(1.6c) are shown in Step 2.

Step 1: Equilibrium bid and waiting time functions. Under this equilibrium conjecture, the expected waiting time function $W^A(c; c_r, c_p, \tilde{c})$ follows from Kleinrock (1967). We concisely reproduce the derivation in our context. For customers $c \in [\underline{c}, c_r) \cup (c_p, \tilde{c}]$, the reasoning is exactly the same as in the proof of Theorem 1. For customers $c \in [c_r, c_p]$, as they submit the same bid and thus expect the same waiting time:

$$W^A(c; c_r, c_p, \tilde{c}) = \frac{1}{\mu - \lambda} - \int_{\underline{c}}^{c_r} \frac{1}{\mu} \Lambda W^e(s; \tilde{c}) dF(s) + \frac{1}{\mu} W^A(c; c_r, c_p, \tilde{c}) \Lambda [F(\tilde{c}) - F(c_p)], \quad c \in [c_r, c_p]. \tag{A.24}$$

Shuffling terms and plugging in the expression of $W^e(s; \tilde{c})$ gives the desired expression in (1.7). Now for cleanness, we use the shorthand notation $b(\cdot)$ for the bid function and $W^A(\cdot)$ for the waiting time function.

We first show that customers prefer bidding $\overline{R}$ to any price β in $[\underline{R}, \overline{R})$. Since the customer who bids $\beta \in [\underline{R}, \overline{R})$ does not trade with any customer whose bids are in $[\underline{R}, \overline{R}]$, bidding

β or $\overline{R}$ will result in the same expected waiting time. However, bidding $\overline{R}$ will result in a higher trading price in selling waiting positions while not affecting the payment made in buying waiting positions; thus the expected utility will be higher. Therefore, bidding $\overline{R}$ is preferred to any bid $b \in [\underline{R}, \overline{R})$. Next, $b(c_r^-)$ must be equal to $\underline{R}$. If $b(c_r^-) < \underline{R}$, then $U(c_r, b(c_r^-)) < U(c_r, b')$ for $b' \in (b(c_r^-), \underline{R})$. This follows a similar argument as above. Thus b' would be a profitable deviation to $b(c_r^-)$. Therefore, $b(c_r^-) = \underline{R}$.

The expected utility of customer c who bids β is

$$U(c,\beta) = \begin{cases} V - cW^A(b^{-1}(\beta)) - \int_{\underline{c}}^{b^{-1}(\beta)} \frac{b(s)}{\mu} \Lambda W^A(s) dF(s) \\ +\frac{\beta}{\mu} W^A(b^{-1}(\beta)) \Lambda [F(\tilde{c}) - F(b^{-1}(\beta))], & \forall \beta \in [b(\underline{c}), \underline{R}) \cup (b(c_p^+), b(\tilde{c})]; \\ V - \frac{c}{\mu[1-\rho F(\tilde{c})+\rho F(c_r)][1-\rho F(\tilde{c})+\rho F(c_p)]} \\ - \int_{\underline{c}}^{c_r} \frac{b(s)}{\mu} \Lambda W^A(s) dF(s) \\ +\frac{\beta}{\mu} \frac{1}{\mu[1-\rho F(\tilde{c})+\rho F(c_r)][1-\rho F(\tilde{c})+\rho F(c_p)]} \Lambda [F(\tilde{c}) - F(c_p)], & \forall \beta = \overline{R}; \end{cases}$$

The symmetric equilibrium strategy $b(c)$ satisfies the first-order condition $\frac{\partial U}{\partial \beta} \mid_{\beta = b(c)} = 0$ for $c \in [\underline{c}, c_r) \cup (c_p, \overline{c}]$. The resulting differential equation is exactly the same as (A.13). The boundary condition for solving the differential equation over $c \in (c_p, \overline{c}]$ is also the same. Therefore, $b^A(c; c_r, c_p, \tilde{c})$ in (1.5) is the same as $b^B(c; \tilde{c})$ in Theorem 1.1 for $c \in (c_p, \overline{c}]$. For solving the differential equation over $[\underline{c}, c_r)$, the boundary condition is $b(c_r^-) = \underline{R}$, and this gives $b^A(c; c_r, c_p, \tilde{c})$ for $c \in [\underline{c}, c_r)$.

Step 2: Using IC and IR in joining to pin down $c_r, c_p, \tilde{c}$. Given $\underline{R}$ and $\overline{R}$, determining the resulting $c_r, c_p, \tilde{c}$ in equilibrium must fall into one of the six distinct cases (see below), depending on the parameter values of $\underline{R}$ and $\overline{R}$.

Case 1: both $\underline{R}$ and $\overline{R}$ are of intermediate values. In this case, $\overline{c} < c_r < c_p < \tilde{c}$. Thus, customer c_r should be indifferent between bidding $\underline{R}$ and $\overline{R}$; i.e, $U(c_r, \underline{R}) = U(c_r, \overline{R})$ (this is captured in (1.6a)). Otherwise, there would be a profitable deviation. For example, if $U(c_r, \underline{R}) < U(c_r, \overline{R})$, customer $c_r - \epsilon$ would deviate from $b(c_r - \epsilon)$ to $\overline{R}$ to gain a higher

expected utility. Similarly, customer c_p should be indifferent between bidding $\overline{R}$ and $b(c_p^+)$; i.e, $U(c_p, \overline{R}) = U(c_p, b(c_p^+))$, captured in (1.6b). Furthermore, following from the same argument of Theorem 1.1, there exists a cutoff type $\tilde{c}$ such that customers with $c \leq \tilde{c}$ join the system. Either $\tilde{c} = \overline{c}$ or $U(\tilde{c}, b^A(\tilde{c}, c_r, c_p, \tilde{c})) = 0$. This is captured in (1.6c) and holds for other cases as well and will not be reiterated. Case 1 is the primary case we focus on. It is later used in Theorem 1.3.

Case 2: both $\underline{R}$ and $\overline{R}$ are so high as not to alter any customer bidding behavior. This can be conceptualized as $c_r = c_p = \tilde{c}$. This case would occur if $\underline{R} > b^B(\tilde{c})$.

Case 3: both $\underline{R}$ and $\overline{R}$ are so low as not to alter any customer bidding behavior. This can be conceptualized as $c_r = c_p = \underline{c}$. This case would occur if $\overline{R} < b^B(\underline{c})$.

Case 4: $\underline{R}$ is intermediate and $\overline{R}$ is high. In this case, $\overline{c} < c_r < c_p = \tilde{c}$. Thus, $U(c_r, \underline{R}) = U(c_r, \overline{R})$ and $U(\tilde{c}, \overline{R}) > U(\tilde{c}, \overline{R} + \epsilon)$.

Case 5: $\underline{R}$ is low and $\overline{R}$ is intermediate. In this case, $\overline{c} = c_r < c_p < \tilde{c}$. Thus, $U(c_p, \overline{R}) = U(c_p, b(c_p^+))$ and $U(\underline{c}, \overline{R}) > U(\underline{c}, \underline{R} - \epsilon)$.

Case 6: $\underline{R}$ is low and $\overline{R}$ is high. In this case, $\overline{c} = c_r < c_p = \tilde{c}$, i.e., all joining customers bid inside $[\underline{R}, \overline{R}]$ and no trade occurs.

Note that (1.6a)-(1.6c) provide a succinct representation subsuming all six cases.

Step 3: Monotonicity. Next we show $b(c)$ is weakly increasing in $c \in [\underline{c}, \tilde{c}]$ and strictly increasing in $c \in [\underline{c}, c_r) \cup (c_p, \tilde{c}]$. To show $b'(c) > 0$ for $c \in [\underline{c}, c_r) \cup (c_p, \tilde{c}]$, this boils down to showing $b(c) > c$ from the differential equation (A.13). This is immediate for $c \in (c_p, \tilde{c}]$ as in the proof of Theorem 1.1. This would hold for $c \in [\underline{c}, c_r)$ if $K_2 > 0$, or $\underline{R} > c_r$. In addition, we need to show $b(c_p^-) \leq b(c_p^+)$, or $\overline{R} \leq b(c_p^+)$ (since $b(c_p^-) = \overline{R}$). We establish monotonicity of the bid function under the context of Case 1 in Step 2, and other cases can be similarly handled.

(i) $c_p < \overline{R} \leq b(c_p^+)$.

$$U(c_p, b(c_p^+)) = V - c_p W^A(c_p^+) - \int_{\underline{c}}^{c_r} \frac{b(s)}{\mu} \Lambda W^A(s) dF(s) - \frac{\overline{R}}{\mu} \Lambda W^A(c_p)[F(c_p) - F(c_r)] + \frac{b(c_p^+)}{\mu} W^A(c_p^+) \Lambda [F(\tilde{c}) - F(c_p)].$$

$$U(c_p, \overline{R}) = V - c_p W^A(c_p) - \int_{\underline{c}}^{c_r} \frac{b(s)}{\mu} \Lambda W^A(s) dF(s) + \frac{\overline{R}}{\mu} W^A(c_p) \Lambda [F(\tilde{c}) - F(c_p)].$$

By Step 2, $U(c_p, b(c_p^+)) = U(c_p, \overline{R})$. This simplifies to

$$\begin{aligned} & c_p W^A(c_p) - c_p W^A(c_p^+) + \rho W^A(c_p^+) b(c_p^+) [F(\tilde{c}) - F(c_p)] \\ &= \rho \overline{R} W^A(c_p) [F(\tilde{c}) - F(c_p)] + \rho \overline{R} [F(c_p) - F(c_r)] W^A(c_p). \end{aligned}$$

Since $b(c_p^+) > c_p$ and $W^A(c_p) \geq W^A(c_p^+)$,

$$\begin{aligned} LHS &> c_p \left\{ W^A(c_p) - W^A(c_p^+) + \rho W^A(c_p^+) [F(\tilde{c}) - F(c_p)] \right\}, \\ LHS &\leq b(c_p^+) \left\{ W^A(c_p) - W^A(c_p^+) + \rho W^A(c_p^+) [F(\tilde{c}) - F(c_p)] \right\}. \end{aligned}$$

To show $c_p < \overline{R} \leq b(c_p^+)$, it suffices to show

$$\rho W^A(c_p)[F(\tilde{c}) - F(c_p)] + \rho [F(c_p) - F(c_r)] W^A(c_p) = W^A(c_p) - W^A(c_p^+) + \rho W^A(c_p^+)[F(\tilde{c}) - F(c_p)].$$

By shuffling terms, this is equivalent to showing

$$W^A(c_p^+) \left[1 - \rho(F(\tilde{c}) - F(c_p)) \right] = W^A(c_p) \left[1 - \rho(F(\tilde{c}) - F(c_r)) \right].$$

Since $W^A(c_p^+) \left[1 - \rho(F(\tilde{c}) - F(c_p)) \right] = \frac{1}{\mu[1 - \rho(F(\tilde{c}) - F(c_p))]}$ and $W^A(c_p) \left[1 - \rho(F(\tilde{c}) - F(c_r)) \right] = \frac{1}{\mu[1 - \rho(F(\tilde{c}) - F(c_p))]}$, this indeed holds and therefore $c_p < \overline{R} < b(c_p^+)$.

(ii) $\underline{R} > c_r$.

$$U(c_p, \underline{R}) = V - c_r W^A(c_r^-) - \int_{\underline{c}}^{c_r} \frac{b(s)}{\mu} \Lambda W^A(s) dF(s) + \frac{\underline{R}}{\mu} W^A(c_r) \Lambda [F(\tilde{c}) - F(c_r)].$$

$$U(c_r, \overline{R}) = V - c_r W^A(c_r) - \int_{\underline{c}}^{c_r} \frac{b(s)}{\mu} \Lambda W^A(s) dF(s) + \frac{\overline{R}}{\mu} W^A(c_r) \Lambda [F(\tilde{c}) - F(c_p)].$$

By Step 2, $U(c_r, \underline{R})) = U(c_r, \overline{R})$. This simplifies to

$$c_r W^A(c_r^-) - c_r W^A(c_r) + \rho \overline{R} W^A(c_r)[F(\tilde{c}) - F(c_p)] = \rho W^A(c_r^-) \underline{R}[F(\tilde{c}) - F(c_r)].$$

Since $\overline{R} > c_p \geq c_r$,

$$c_r W^A(c_r^-) - c_r W^A(c_r) + \rho c_r W^A(c_r)[F(\tilde{c}) - F(c_p)] < \rho W^A(c_r^-) \underline{R}[F(\tilde{c}) - F(c_r)].$$

To show $c_r < \underline{R}$, it suffices to show

$$W^A(c_r^-) - W^A(c_r) + \rho W^A(c_r)[F(\tilde{c}) - F(c_p)] = \rho W^A(c_r^-)[F(\tilde{c}) - F(c_r)].$$

Or equivalently,

$$W^A(c_r)[1 - \rho(F(\tilde{c}) - F(c_p))] = W^A(c_r^-)[1 - \rho(F(\tilde{c}) - F(c_r))].$$

Since $W^A(c_r)[1 - \rho(F(\tilde{c}) - F(c_p))] = \frac{1}{\mu[1-\rho(F(\tilde{c})-F(c_r))]}$ and $W^A(c_r^-)[1 - \rho(F(\tilde{c}) - F(c_r))] = \frac{1}{\mu[1-\rho(F(\tilde{c})-F(c_r))]}$, this indeed holds, and therefore $c_r < \underline{R}$.

Hence we prove the bidding function is indeed weakly increasing as conjectured.

Step 4: IR in trading. As in the proof of Theorem 1.1, we need to check if $\vartheta(c) = U(c, b(c)) - [V - \frac{c}{\mu - \lambda}] \geq 0$ for all $c \in [\underline{c}, \tilde{c}]$. Since $\vartheta(c)$ is weakly convex (this can be similarly proved as in Theorem 1.1), there are three cases.

Case 1: $\arg\min \vartheta(c) \not\subseteq [c_r, c_p]$. Then the analysis in the proof of Theorem 1.1 remains

valid.

Case 2: $c_r = \arg\min \vartheta(c)$. This is true when $W^A(c_r^-) \geq \overline{W}(\tilde{c})$ and $W^A(c_r) \leq \overline{W}(\tilde{c})$. This implies $c_r \leq c_m(\tilde{c})$. We shall show $U(c_r, b(c_r)) \geq V - c_r\overline{W}(\tilde{c})$. Since

$$U(c_r, b(c_r)) = V - c_r W^A(c_r^-) - \rho \int_{\underline{c}}^{c_r} b(c) W^A(c) dc + \rho \underline{R} W^A(c_r^-)[F(\tilde{c}) - F(c_r)],$$

it is equivalent to showing

$$c_r \left[W^A(c_r^-) - \overline{W}(\tilde{c}) \right] \leq \rho \underline{R} W^A(c_r^-)[F(\tilde{c}) - F(c_r)] - \rho \int_{\underline{c}}^{c_r} b(c) W^A(c) dc. \tag{A.25}$$

Since $b(c) \leq \underline{R}$ for $c \in [\underline{c}, c_r)$, therefore,

$$\rho \int_{\underline{c}}^{c_r} b(c) W^A(c) dc \leq \underline{R} \int_{\underline{c}}^{c_r} \rho W^A(c) dc = \underline{R} \left[\overline{W}(\tilde{c}) - \frac{1}{\mu(1 - \rho F(\tilde{c}) + \rho F(c_r))} \right].$$

Therefore,

$$\begin{aligned} &\rho \underline{R} W^A(c_r^-)[F(\tilde{c}) - F(c_r)] - \rho \int_{\underline{c}}^{c_r} b(c) W^A(c; \tilde{c}) dc \\ &\geq \underline{R} \left[\frac{\rho[F(\tilde{c}) - F(c_r)]}{\mu(1 - \rho F(\tilde{c}) + \rho F(c_r))^2} - \overline{W}(\tilde{c}) + \frac{1}{\mu(1 - \rho F(\tilde{c}) + \rho F(c_r))} \right]. \end{aligned}$$

Since $c_r < \underline{R}$ and by simple algebra,

$$\frac{\rho[F(\tilde{c}) - F(c_r)]}{\mu(1 - \rho F(\tilde{c}) + \rho F(c_r))^2} + \frac{1}{\mu(1 - \rho F(\tilde{c}) + \rho F(c_r))} = \frac{1}{\mu(1 - \rho F(\tilde{c}) + \rho F(c_r))^2},$$

thus,

$$\begin{aligned} c_r \left[W^A(c_r^-) - \overline{W}(\tilde{c}) \right] &\leq \underline{R} \left[\frac{\rho[F(\tilde{c}) - F(c_r)]}{\mu(1 - \rho F(\tilde{c}) + \rho F(c_r))^2} - \overline{W}(\tilde{c}) + \frac{1}{\mu(1 - \rho F(\tilde{c}) + \rho F(c_r))} \right] \\ &\leq \rho \underline{R} W^A(c_r^-)[F(\tilde{c}) - F(c_r)] - \rho \int_{\underline{c}}^{c_r} b(c) W^A(c) dc. \end{aligned}$$

This proves (A.25).

Case 3: $c_p = \arg\min \vartheta(c)$. $U(c_p, b(c_p)) \geq V - c_p\overline{W}(\tilde{c})$ can be shown similarly as in Case 2.

Note that if $W^A(c_r) = \overline{W}(\tilde{c})$, $\arg\min \vartheta(c)$ is the set $[c_r, c_p]$, but this is subsumed in Case 2 and 3. □

Proof of Theorem 1.3. (1.9a)-(1.9c) restates the optimal direct mechanism in Problem A.3 derived in Appendix A.1. Given the optimal $(c_r, c_p, \tilde{c})$, we find the parameters $(H, \underline{R}, \overline{R})$ according to (1.8a)-(1.8c). We shall show this auction implements the intermediary's optimal direct mechanism and thus raises the maximum revenue for the intermediary by the revenue equivalence theorem. Theorem 1.2 gives the equilibrium structure of the $(H, \underline{R}, \overline{R})$ auction. First, from the expected waiting time in Theorem 1.2-(iii) and the condition that $c_r, c_p, \tilde{c}$ satisfy (1.9b), we verify that the expected waiting time of the auction matches that of the optimal direct mechanism in (A.7) in Lemma A.3-(i). Second, in the equilibrium of the auction, no customers benefit from deviating from their own bid. This corresponds to the IC constraints in the direct mechanism. Specifically, the equilibrium conditions (1.6a)-(1.6b) require each of c_r and c_p to be indifferent between bidding two different prices: $U(c_r, b(c_r^-)) = U(c_r, b(c_r^+))$ and $U(c_p, b(c_p^-)) = U(c_p, b(c_p^+))$. Writing them out gives

$$\frac{c_r}{\mu-\lambda} - \frac{1}{\mu-\lambda}\rho\overline{R}[F(\tilde{c}) - F(c_p)] = c_r W(c_r) - \rho W(c_r)\underline{R}[F(\tilde{c}) - F(c_r)] \tag{A.26}$$

$$\frac{c_p}{\mu-\lambda} - \frac{1}{\mu-\lambda}\rho\overline{R}[F(\tilde{c}) - F(c_p)] = c_p W(c_p) - \rho W(c_p) b(c_p)[F(\tilde{c}) - F(c_p)]) + \frac{\rho\overline{R}[F(c_p) - F(c_r)]}{\mu-\lambda} \tag{A.27}$$

Note that here we have applied the substitution $\frac{1}{\mu-\lambda} = \frac{1}{\mu(1-\rho F(\tilde{c})+\rho F(c_r))(1-\rho F(\tilde{c})+\rho F(c_p))}$ required by (1.9b). Shuffling terms of (A.26) and (A.27) gives (1.8b)-(1.8c) that express $\underline{R}^*$ and $\overline{R}^*$ in terms of $c_r^*, c_p^*, \tilde{c}^*$. It is also easy to verify $\underline{R}^* \leq \overline{R}^*$ by (1.8b)-(1.8c). Finally, (1.8a) ensures that the revenues in the auction and in the optimal direct mechanism are equal. Therefore, the $(H^*, \underline{R}^*, \overline{R}^*)$ auction implements the intermediary's optimal direct mechanism, They are outcome equivalent. By Lemma A.5, $U(c) = V - c\overline{W}(\tilde{c})$ for $c \in$

$[c_r, c_p]$ in the direct mechanism. Customers in $[c_r, c_p]$ must expected the same utility in the $(H^*, \underline{R}^*, \overline{R}^*)$ auction, and this implies (1.10). □

Proof of Theorem 1.4. If (1.9c) is not binding, then Lemma A.3-(iii) indicates $f_r(c_r^*) = f_p(c_p^*; \tilde{c}^*)$. It follows that $c_r^* < c_p^*$.

Now we show $c_r^* < c_p^*$ also holds if (1.9c) is binding, which we denote by $G(c_p, \tilde{c}) = 0$. Write $c_p(\tilde{c})$ from (1.9c) and $c_r(\tilde{c})$ from (1.9b).

$$\frac{\partial G}{\partial c_p} = W^e(c_p; \tilde{c}) - \overline{W}(\tilde{c}), \qquad \frac{\partial G}{\partial \tilde{c}} = -\int_{c_p}^{\tilde{c}} \frac{\partial W^e(c; \tilde{c})}{\partial \tilde{c}} dc - \frac{1}{\mu} - c_p \frac{\partial \overline{W}(\tilde{c})}{\partial \tilde{c}}.$$

By the implicit function theorem,

$$\frac{dc_p}{d\tilde{c}} = -\frac{\frac{\partial G}{\partial \tilde{c}}}{\frac{\partial G}{\partial c_p}} = \frac{\int_{c_p}^{\tilde{c}} \frac{\partial W^e(c;\tilde{c})}{\partial \tilde{c}} dc + \frac{1}{\mu} + c_p \frac{\partial \overline{W}(\tilde{c})}{\partial \tilde{c}}}{W^e(c_p; \tilde{c}) - \overline{W}(\tilde{c})}. \tag{A.28}$$

(1.9b) can be written as

$$Y \triangleq \int_{c_r}^{c_p} W^e(c; \tilde{c}) dF(c) - (F(c_p) - F(c_r))\overline{W}(\tilde{c}) = 0,$$

$$\frac{\partial Y}{\partial c_r} = -W^e(c_r; \tilde{c}) f(c_r) + f(c_r)\overline{W}(\tilde{c}), \qquad \frac{\partial Y}{\partial c_p} = W^e(c_p; \tilde{c}) f(c_p) - f(c_p)\overline{W}(\tilde{c}),$$

$$\frac{\partial Y}{\partial \tilde{c}} = \int_{c_r}^{c_p} \frac{\partial W^e(c; \tilde{c})}{\partial \tilde{c}} dF(c) - (F(c_p) - F(c_r)) \frac{\partial \overline{W}(\tilde{c})}{\partial \tilde{c}}.$$

By the implicit function theorem,

$$\frac{\partial c_r}{\partial c_p} = -\frac{\frac{\partial Y}{\partial c_p}}{\frac{\partial Y}{\partial c_r}} = \frac{f(c_p)\left[W^e(c_p; \tilde{c}) - \overline{W}(\tilde{c})\right]}{f(c_r)\left[W^e(c_r; \tilde{c}) - \overline{W}(\tilde{c})\right]}. \tag{A.29}$$

$$\frac{\partial c_r}{\partial \tilde{c}} = -\frac{\frac{\partial Y}{\partial \tilde{c}}}{\frac{\partial Y}{\partial c_r}} = \frac{\int_{c_r}^{c_p} \frac{\partial W^e(c;\tilde{c})}{\partial \tilde{c}} dF(c) - (F(c_p) - F(c_r)) \frac{\partial \overline{W}(\tilde{c})}{\partial \tilde{c}}}{f(c_r)\left[W^e(c_r; \tilde{c}) - \overline{W}(\tilde{c})\right]}. \tag{A.30}$$

By Leibniz integral rule,

$$\begin{aligned}
&\frac{d\Pi(c_r(\tilde{c}),c_p(\tilde{c}),\tilde{c})}{d\tilde{c}} \\
=&-\Lambda\left\{\begin{array}{l}
\int_{\underline{c}}^{c_r(\tilde{c})}\frac{\partial[W^e(c;\tilde{c})-\overline{W}(\tilde{c})]}{\partial\tilde{c}}f_r(c)dF(c)+\left[W^e(c_r;\tilde{c})-\overline{W}(\tilde{c})\right]f_r(c_r)f(c_r)\frac{dc_r}{d\tilde{c}} \\
\quad+\int_{c_p(\tilde{c})}^{\tilde{c}}\left[\frac{\partial[W^e(c;\tilde{c})-\overline{W}(\tilde{c})]}{\partial\tilde{c}}f_p(c;\tilde{c})+\left[W^e(c;\tilde{c})-\overline{W}(\tilde{c})\right]\frac{\partial f_p(c;\tilde{c})}{\partial\tilde{c}}\right]dF(c) \\
+\left(W^e(\tilde{c};\tilde{c})-\overline{W}(\tilde{c})\right)f_p(\tilde{c};\tilde{c})f(\tilde{c})-\left[W^e(c_p;\tilde{c})-\overline{W}(\tilde{c})\right]f_p(c_p;\tilde{c})f(c_p)\frac{dc_p}{d\tilde{c}}
\end{array}\right\}.
\end{aligned}$$

From (A.28),

$$\begin{aligned}
&\left[W^e(c_p;\tilde{c})-\overline{W}(\tilde{c})\right]f_p(c_p;\tilde{c})f(c_p)\frac{dc_p}{d\tilde{c}} \\
=&\left[W^e(c_p;\tilde{c})-\overline{W}(\tilde{c})\right]f_p(c_p;\tilde{c})f(c_p)\frac{\int_{c_p}^{\tilde{c}}\frac{\partial W^e(c;\tilde{c})}{\partial\tilde{c}}dc+\frac{1}{\mu}+c_p\frac{\partial\overline{W}(\tilde{c})}{\partial\tilde{c}}}{W^e(c_p;\tilde{c})-\overline{W}(\tilde{c})} \\
=&f_p(c_p;\tilde{c})f(c_p)\left[\int_{c_p}^{\tilde{c}}\frac{\partial W^e(c;\tilde{c})}{\partial\tilde{c}}dc+\frac{1}{\mu}+c_p\frac{\partial\overline{W}(\tilde{c})}{\partial\tilde{c}}\right].
\end{aligned}$$

By the chain's rule, $\frac{dc_r}{d\tilde{c}}=\frac{\partial c_r}{\partial\tilde{c}}+\frac{\partial c_r}{\partial c_p}\frac{dc_p}{d\tilde{c}}$. Therefore,

$$\left[W^e(c_r;\tilde{c})-\overline{W}(\tilde{c})\right]f_r(c_r)f(c_r)\frac{dc_r}{d\tilde{c}}=\left[W^e(c_r;\tilde{c})-\overline{W}(\tilde{c})\right]f_r(c_r)f(c_r)\left[\frac{\partial c_r}{\partial\tilde{c}}+\frac{\partial c_r}{\partial c_p}\frac{dc_p}{d\tilde{c}}\right].$$

To show $c_r^*<c_p^*$, it suffices to show that if $c_r(\tilde{c})=c_p(\tilde{c})$,

$$\frac{d\Pi(c_r(\tilde{c}),c_p(\tilde{c}),\tilde{c})}{d\tilde{c}}<0. \tag{A.31}$$

Let $c_r(\tilde{c})=c_p(\tilde{c})\triangleq c_m(\tilde{c})$. Thus, $W^e(c_m(\tilde{c});\tilde{c})=\overline{W}(\tilde{c})$ from (1.9b) and $\frac{\partial c_r}{\partial c_p}=1$ from (A.29).

Hence from (1.9b) and (A.29) and (A.30),

$$
\begin{aligned}
&\left[W^e(c_r;\tilde{c}) - \overline{W}(\tilde{c})\right] f_r(c_r) f(c_r) \frac{dc_r}{d\tilde{c}} \\
&= \left[W^e(c_r;\tilde{c}) - \overline{W}(\tilde{c})\right] f_r(c_r) f(c_r) \left[\frac{dc_p}{d\tilde{c}} + \frac{\int_{c_r}^{c_p} \frac{\partial W^e(c;\tilde{c})}{\partial \tilde{c}} dF(c) - (F(c_p) - F(c_r)) \frac{\partial \overline{W}(\tilde{c})}{\partial \tilde{c}}}{f(c_r)\left[W^e(c_r;\tilde{c}) - \overline{W}(\tilde{c})\right]} \right] \\
&= \left[W^e(c_m(\tilde{c});\tilde{c}) - \overline{W}(\tilde{c})\right] f_r(c_m(\tilde{c})) f(c_m(\tilde{c})) \frac{dc_p}{d\tilde{c}} \\
&= \left[W^e(c_m(\tilde{c});\tilde{c}) - \overline{W}(\tilde{c})\right] f_r(c_m(\tilde{c})) f(c_m(\tilde{c})) \frac{\int_{c_m(\tilde{c})}^{\tilde{c}} \frac{\partial W^e(c;\tilde{c})}{\partial \tilde{c}} dc + \frac{1}{\mu} + c_m(\tilde{c}) \frac{\partial \overline{W}(\tilde{c})}{\partial \tilde{c}}}{W^e(c_m(\tilde{c});\tilde{c}) - \overline{W}(\tilde{c})} \\
&= f_r(c_m(\tilde{c})) f(c_m(\tilde{c})) \left[\int_{c_m(\tilde{c})}^{\tilde{c}} \frac{\partial W^e(c;\tilde{c})}{\partial \tilde{c}} dc + \frac{1}{\mu} + c_m(\tilde{c}) \frac{\partial \overline{W}(\tilde{c})}{\partial \tilde{c}} \right].
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\frac{d\Pi(c_r(\tilde{c}), c_p(\tilde{c}), \tilde{c})}{d\tilde{c}} \\
&= -\Lambda \left\{ \begin{array}{c}
\int_{\underline{c}}^{c_m(\tilde{c})} \frac{\partial\left[W^e(c;\tilde{c}) - \overline{W}(\tilde{c})\right]}{\partial \tilde{c}} f_r(c) dF(c) + f_r(c_m(\tilde{c})) f(c_m(\tilde{c})) \left[\int_{c_m(\tilde{c})}^{\tilde{c}} \frac{\partial W^e(c;\tilde{c})}{\partial \tilde{c}} dc + \frac{1}{\mu} + c_m(\tilde{c}) \frac{\partial \overline{W}(\tilde{c})}{\partial \tilde{c}} \right] \\
+ \int_{c_m(\tilde{c})}^{\tilde{c}} \left[\frac{\partial\left[W^e(c;\tilde{c}) - \overline{W}(\tilde{c})\right]}{\partial \tilde{c}} f_p(c;\tilde{c}) + \left[W^e(c;\tilde{c}) - \overline{W}(\tilde{c})\right] \frac{\partial f_p(c;\tilde{c})}{\partial \tilde{c}} \right] dF(c) \\
+ \left(W^e(\tilde{c};\tilde{c}) - \overline{W}(\tilde{c})\right) f_p(\tilde{c};\tilde{c}) f(\tilde{c}) - f_p(c_m(\tilde{c});\tilde{c}) f(c_m(\tilde{c})) \left[\int_{c_m(\tilde{c})}^{\tilde{c}} \frac{\partial W^e(c;\tilde{c})}{\partial \tilde{c}} dc + \frac{1}{\mu} + c_m(\tilde{c}) \frac{\partial \overline{W}(\tilde{c})}{\partial \tilde{c}} \right]
\end{array} \right\} \\
&= -\Lambda \left\{ \begin{array}{c}
\int_{\underline{c}}^{c_m(\tilde{c})} \frac{\partial\left[W^e(c;\tilde{c}) - \overline{W}(\tilde{c})\right]}{\partial \tilde{c}} f_r(c) dF(c) \\
+ \left[f_r(c_m) - f_p(c_m;\tilde{c})\right] f(c_m) \left[\int_{c_m}^{\tilde{c}} \frac{\partial W^e(c;\tilde{c})}{\partial \tilde{c}} dc + \frac{1}{\mu} + c_m(\tilde{c}) \frac{\partial \overline{W}(\tilde{c})}{\partial \tilde{c}} \right] \\
+ \int_{c_m(\tilde{c})}^{\tilde{c}} \left[\frac{\partial\left[W^e(c;\tilde{c}) - \overline{W}(\tilde{c})\right]}{\partial \tilde{c}} f_p(c;\tilde{c}) + \left[W^e(c;\tilde{c}) - \overline{W}(\tilde{c})\right] \frac{\partial f_p(c;\tilde{c})}{\partial \tilde{c}} \right] dF(c) \\
+ \left(W^e(\tilde{c};\tilde{c}) - \overline{W}(\tilde{c})\right) f_p(\tilde{c};\tilde{c}) f(\tilde{c})
\end{array} \right\} \\
&= -\Lambda \left\{ \begin{array}{c}
\int_{\underline{c}}^{c_m(\tilde{c})} \frac{\partial\left[W^e(c;\tilde{c}) - \overline{W}(\tilde{c})\right]}{\partial \tilde{c}} f_r(c) dF(c) \\
+ F(\tilde{c}) \left[\int_{c_m(\tilde{c})}^{\tilde{c}} \frac{\partial W^e(c;\tilde{c})}{\partial \tilde{c}} dc + \frac{1}{\mu} + c_m(\tilde{c}) \frac{\partial \overline{W}(\tilde{c})}{\partial \tilde{c}} \right] \\
+ \int_{c_m(\tilde{c})}^{\tilde{c}} \left[\frac{\partial\left[W^e(c;\tilde{c}) - \overline{W}(\tilde{c})\right]}{\partial \tilde{c}} f_p(c;\tilde{c}) + \left[\overline{W}(\tilde{c}) - W^e(c;\tilde{c})\right] \frac{f(\tilde{c})}{f(c)} \right] dF(c) \\
+ \left(\frac{1}{\mu} - \overline{W}(\tilde{c})\right) \tilde{c} f(\tilde{c})
\end{array} \right\} \\
&\triangleq -\Lambda \Psi(\tilde{c}).
\end{aligned}
$$

Thus to show (A.31), it suffices to show $\Psi(\tilde{c}) > 0$. Note that $\int_{\underline{c}}^{\tilde{c}} f_r(c)dF(c) = \tilde{c}F(\tilde{c})$.

$$\begin{aligned}
\Psi(\tilde{c}) = &\int_{\underline{c}}^{\tilde{c}} \frac{\partial \left[W^e(c;\tilde{c}) - \overline{W}(\tilde{c})\right]}{\partial \tilde{c}} f_r(c)dF(c) \\
&+ F(\tilde{c}) \left[\int_{c_m(\tilde{c})}^{\tilde{c}} \frac{\partial W^e(c;\tilde{c})}{\partial \tilde{c}} dc + \frac{1}{\mu} + c_m(\tilde{c}) \frac{\partial \overline{W}(\tilde{c})}{\partial \tilde{c}} - \int_{c_m(\tilde{c})}^{\tilde{c}} \frac{\partial \left[W^e(c;\tilde{c}) - \overline{W}(\tilde{c})\right]}{\partial \tilde{c}} dc\right] \\
&+ f(\tilde{c}) \int_{c_m(\tilde{c})}^{\tilde{c}} \left[\overline{W}(\tilde{c}) - W^e(c;\tilde{c})\right] dc + \left(\frac{1}{\mu} - \overline{W}(\tilde{c})\right) \tilde{c} f(\tilde{c}) \\
= &\int_{\underline{c}}^{\tilde{c}} \frac{\partial W^e(c;\tilde{c})}{\partial \tilde{c}} f_r(c)dF(c) - \frac{\partial \overline{W}(\tilde{c})}{\partial \tilde{c}} \tilde{c}F(\tilde{c}) + F(\tilde{c}) \left[\frac{1}{\mu} + \tilde{c} \frac{\partial \overline{W}(\tilde{c})}{\partial \tilde{c}}\right] \\
&- f(\tilde{c}) \overline{W}(\tilde{c}) c_m(\tilde{c}) - f(\tilde{c}) \int_{c_m(\tilde{c})}^{\tilde{c}} W^e(c;\tilde{c})dc + \frac{\tilde{c}f(\tilde{c})}{\mu} \\
= &\int_{\underline{c}}^{\tilde{c}} \frac{\partial W^e(c;\tilde{c})}{\partial \tilde{c}} (cf(c) + F(c))dc + \frac{F(\tilde{c}) + \tilde{c}f(\tilde{c})}{\mu} - f(\tilde{c}) \overline{W}(\tilde{c}) c_m(\tilde{c}) - f(\tilde{c}) \int_{c_m(\tilde{c})}^{\tilde{c}} W^e(c;\tilde{c})dc.
\end{aligned}$$

Note that $W^e(c_m(\tilde{c});\tilde{c}) = \overline{W}(\tilde{c})$. By integration by parts,

$$\begin{aligned}
&\frac{F(\tilde{c}) + \tilde{c}f(\tilde{c})}{\mu} - f(\tilde{c}) \overline{W}(\tilde{c}) c_m(\tilde{c}) - f(\tilde{c}) \int_{c_m(\tilde{c})}^{\tilde{c}} W^e(c;\tilde{c})dc \\
&= \frac{F(\tilde{c}) + \tilde{c}f(\tilde{c})}{\mu} - f(\tilde{c}) \overline{W}(\tilde{c}) c_m(\tilde{c}) - f(\tilde{c}) \left[\tilde{c}/\mu - \overline{W}(\tilde{c}) c_m(\tilde{c}) - \int_{c_m(\tilde{c})}^{\tilde{c}} c \frac{\partial W^e(c;\tilde{c})}{\partial c} dc\right] \\
&= \frac{F(\tilde{c})}{\mu} - f(\tilde{c}) \int_{c_m(\tilde{c})}^{\tilde{c}} c \frac{-\partial W^e(c;\tilde{c})}{\partial c} dc.
\end{aligned}$$

Therefore,

$$\Psi(\tilde{c}) = \int_{\underline{c}}^{c_m(\tilde{c})} \frac{\partial W^e(c;\tilde{c})}{\partial \tilde{c}} f_r(c)dF(c) + \int_{c_m(\tilde{c})}^{\tilde{c}} \left[\frac{\partial W^e(c;\tilde{c})}{\partial \tilde{c}} (cf(c) + F(c)) - f(\tilde{c}) c \frac{-\partial W^e(c;\tilde{c})}{\partial c}\right] dc + \frac{F(\tilde{c})}{\mu}.$$

Since

$$\frac{\partial W^e(c;\tilde{c})}{\partial c} = \frac{-2\rho f(c)}{\mu(1 - \rho F(\tilde{c}) + \rho F(c))^3}, \qquad \frac{\partial W^e(c;\tilde{c})}{\partial \tilde{c}} = \frac{2\rho f(\tilde{c})}{\mu(1 - \rho F(\tilde{c}) + \rho F(c))^3} > 0,$$

therefore,

$$\begin{aligned}\Psi(\tilde{c}) &= \int_{\underline{c}}^{c_m(\tilde{c})} \frac{\partial W^e(c;\tilde{c})}{\partial \tilde{c}} f_r(c) dF(c) + \frac{F(\tilde{c})}{\mu} \\ &\quad + \int_{c_m(\tilde{c})}^{\tilde{c}} \left[\frac{2\rho f(\tilde{c})}{\mu(1-\rho F(\tilde{c})+\rho F(c))^3}(cf(c)+F(c)) - f(\tilde{c})c\frac{2\rho f(c)}{\mu(1-\rho F(\tilde{c})+\rho F(c))^3} \right] dc \\ &= \int_{\underline{c}}^{c_m(\tilde{c})} \frac{\partial W^e(c;\tilde{c})}{\partial \tilde{c}} f_r(c) dF(c) + \frac{F(\tilde{c})}{\mu} + \int_{c_m(\tilde{c})}^{\tilde{c}} \frac{2\rho f(\tilde{c})F(c)}{\mu(1-\rho F(\tilde{c})+\rho F(c))^3} dc \\ &= \int_{\underline{c}}^{c_m(\tilde{c})} \frac{\partial W^e(c;\tilde{c})}{\partial \tilde{c}} f_r(c) dF(c) + \int_{c_m(\tilde{c})}^{\tilde{c}} \frac{\partial W^e(c;\tilde{c})}{\partial \tilde{c}} F(c) dc + \frac{F(\tilde{c})}{\mu} \\ &>0. \end{aligned}$$

□

Proof of Proposition 1.3. Recall that we show that given $\tilde{c}$, $\vartheta(c) = U(c, b^B(c;\tilde{c})) - [V - c\overline{W}(\tilde{c})]$ attain its minimum at $c_m(\tilde{c})$ ($W^e(c_m(\tilde{c});\tilde{c}) = \overline{W}(\tilde{c})$). Therefore, $\overline{H}$ satisfies $\vartheta(c_m) - \overline{H} = 0$, which implies that minimum of $\vartheta(c) - \overline{H}$ is equal to zero. Hence, all joining customers still voluntarily trade with customer $c_m(\tilde{c})$ being indifferent between FIFO and trading. Shuffling terms of $\vartheta(c_m) - \overline{H} = 0$ gives $U(c_m(\tilde{c}), b^B(c_m(\tilde{c});\tilde{c})) - \overline{H} = V - c_m(\tilde{c})\overline{W}(\tilde{c})$. From the proof of Theorem 1.1, $U(\tilde{c}; b^B(\tilde{c};\tilde{c})) = U(c_m(\tilde{c}), b^B(c_m(\tilde{c});\tilde{c})) - \int_{c_m(\tilde{c})}^{\tilde{c}} W^e(c;\tilde{c})dc$. It follows that the expected utility of customer $\tilde{c}$ in the $\overline{H}$ auction is $U(\tilde{c}; b^B(\tilde{c};\tilde{c})) - \overline{H} = V - c_m(\tilde{c})\overline{W}(\tilde{c}) - \int_{c_m(\tilde{c})}^{\tilde{c}} W^e(c;\tilde{c})dc$. Define

$$\Upsilon(\tilde{c}, x) \triangleq V - \int_x^{\tilde{c}} W^e(c;\tilde{c})dc - x\overline{W}(\tilde{c}).$$

Thus, in the optimal auction, either $\Upsilon(\tilde{c}^*, c_p^*) = 0$ or $\tilde{c}^* = \overline{c}$; in a FIFO queue, either $\Upsilon(\tilde{c}^{\text{FIFO}}, \tilde{c}^{\text{FIFO}}) = 0$ or $\tilde{c}^{\text{FIFO}} = \overline{c}$; in the $\overline{H}$ auction, either $\Upsilon(\tilde{c}^{\overline{H}}, c_m(\tilde{c}^{\overline{H}})) = 0$ or $\tilde{c}^{\overline{H}} = \overline{c}$. Since $\frac{\partial \Upsilon(\tilde{c},x)}{\partial x} = W^e(x;\tilde{c}) - \overline{W}(\tilde{c}) < 0$ if $x > c_m(\tilde{c})$, and $c_m(\tilde{c}) < c_p(\tilde{c}) < \tilde{c}$ for any given $\tilde{c}$, it follows that

$$\Upsilon(\tilde{c}, c_m(\tilde{c})) > \Upsilon(\tilde{c}, c_p(\tilde{c})) > \Upsilon(\tilde{c}, \tilde{c}). \tag{A.32}$$

Since

$$\begin{aligned}\frac{d\Upsilon(\tilde{c}, c_m(\tilde{c}))}{d\tilde{c}} &= -\left[\frac{1}{\mu} - \overline{W}(\tilde{c})\frac{dc_m(\tilde{c})}{d\tilde{c}}\right] - \int_{c_m(\tilde{c})}^{\tilde{c}} \frac{\partial W^e(c;\tilde{c})}{\partial \tilde{c}} dc - \overline{W}(\tilde{c})\frac{dc_m(\tilde{c})}{d\tilde{c}} - c_m(\tilde{c})\frac{\partial W^e(c;\tilde{c})}{\partial \tilde{c}} \\ &= -\frac{1}{\mu} - \int_{c_m(\tilde{c})}^{\tilde{c}} \frac{\partial W^e(c;\tilde{c})}{\partial \tilde{c}} dc - c_m(\tilde{c})\frac{\partial W^e(c;\tilde{c})}{\partial \tilde{c}} < 0 \text{ and} \\ \frac{d\Upsilon(\tilde{c},\tilde{c})}{d\tilde{c}} &= -\tilde{c}\frac{\partial \overline{W}(\tilde{c})}{\partial \tilde{c}} - \overline{W}(\tilde{c}) < 0,\end{aligned} \tag{A.33}$$

it follows that if $\tilde{c}^* < \overline{c}$ ($\Upsilon(\tilde{c}^*, c_p^*) = 0$), then from (A.32), $\Upsilon(\tilde{c}^*, c_m(\tilde{c}^*)) > 0$ and $\Upsilon(\tilde{c}^*, \tilde{c}^*) < 0$; thus, by (A.33), $\tilde{c}^{\text{FIFO}} < \tilde{c}^* < \tilde{c}^{\overline{H}}$. Likewise, one can show that if $\tilde{c}^* = \overline{c}$, then $\tilde{c}^{\text{FIFO}} \leq \tilde{c}^* = \tilde{c}^{\overline{H}}$. Since the ordering of λ^{FIFO}, λ^*, and $\lambda^{\overline{H}}$ is the same as the ordering of c^{FIFO}, c^*, and $c^{\overline{H}}$, therefore the result follows. □

A.5 Proofs of Lemmas A.1-A.5

Proof of Lemma A.1. Refer to Problem A.1. The IC constraints (A.1d) require

$$U(c) \geq q(c')[V - cW(c') - P(c')] = U(c') + q(c')W(c')(c' - c).$$

Similarly, $U(c') \geq U(c) + W(c)q(c)(c - c')$. Therefore, the IC constraints are equivalent to

$$-q(c)W(c)(c - c') \geq U(c) - U(c') \geq -q(c')W(c')(c - c').$$

Therefore, $q(c)W(c)$ is weakly decreasing in c. It follows that $W(c)$ is Riemann integrable, and by the squeeze theorem, $U'(c) = -q(c)W(c)$ almost everywhere. Since $q(c) \in \{0, 1\}$ and $U(c) = 0$ if $q(c) = 0$, the IR constraints $U(c) \geq 0$ $\forall c \in \Xi$ imply that there exists a $\tilde{c}$ such that $q(c) = 1$ if $c \leq \tilde{c}$ and $q(c) = 0$ otherwise. Moreover, if $\tilde{c} < \overline{c}$ then $U(\tilde{c}) = 0$. Therefore, (A.3) and (A.2) follow. Since $U'(c) = -W(c)$ almost everywhere for $c \in [\underline{c}, \tilde{c}]$, it follows that $U(c)$ is decreasing in c and since $W(c)$ is weakly decreasing in $c \in [\underline{c}, \tilde{c}]$, this further implies

$U(c)$ is convex. Finally, for $c \in [\underline{c}, \tilde{c}]$, the IC constraints are

$$V - cW(c) - P(c) \geq V - cW(c') - P(c') \Leftrightarrow P(c') - P(c) \geq -c[W(c') - W(c)].$$

This implies $P(c)$ is weakly increasing. □

Proof of Lemma A.3. Refer to Problem A.2.

Part (i): Since $W(\cdot)$ is nonincreasing, the OA constraints simplify to:

$$\int_c^{\tilde{c}} W(s)dF(s) \geq \frac{F(\tilde{c}) - F(c)}{\mu - \Lambda(F(\tilde{c}) - F(c))} \qquad \forall c \in [c_p, \tilde{c}] \tag{A.34}$$

$$\int_c^{c_r} W(s)dF(s) + \frac{F(c_p) - F(c_r)}{\mu - \Lambda F(\tilde{c})} + \int_{c_p}^{\tilde{c}} W(s)dF(s) \geq \frac{F(\tilde{c}) - F(c)}{\mu - \Lambda(F(\tilde{c}) - F(c))} \qquad \forall c \in [\underline{c}, c_r] \tag{A.35}$$

In addition, $W(c_1) > \frac{1}{\mu - \Lambda F(\tilde{c})} > W(c_2)$ for $c_1 < c_r$ and $c_2 > c_p$.

Case 1: (A.5b) is not binding. Then $p_m^* = 0$. if $p_m^* < 0$, then there exists $\epsilon > 0$ such that $p_m^0 = p_m^* + \epsilon < 0$ and that (A.5b) is still not binding with p_m^0 and that the objective function is increased. Hence, $p_m^* = 0$.

Claim 1: $f_p(c, \tilde{c}) > 0$ for $c \in (c_p, \tilde{c}]$. Proof by contradiction. Suppose there exists $c_1 \in (c_p, \tilde{c}]$ such that $f_p(c_1; \tilde{c}) \leq 0$. Since $\frac{\partial df_p(c;\tilde{c})}{\partial c} > 0$, then $f_p(c) < 0$ and $W^*(c) < \overline{W}(\tilde{c})$ for $c \in (c_p, c_1)$. Then one can construct $W_1(c) = \overline{W}(\tilde{c})$ for $c \in (c_p, c_1)$ and $W_1(c) = W^*(c)$ elsewhere. This $W_1(\cdot)$ strictly increases the objective function (A.5a) over $W^*(c)$. Therefore, $f_p(c, \tilde{c}) > 0$ for $c \in (c_p, \tilde{c}]$.

Claim 2: (A.34) must be binding. Suppose not, then there exists a feasible perturbation that decreases $W^*(c)$ for types $(c_1, c_1 + \epsilon_1] \subset (c_p^*, \tilde{c}]$ and increases $W^*(c)$ for some types $(c_2, c_2 + \epsilon_2] \subset (c_p^*, \tilde{c}]$ with $\epsilon_1, \epsilon_2 > 0$ and $c_2 + \epsilon_2 \leq c_1$. This feasible perturbation strictly increases the objective function (A.5a) since $f_p(c, \tilde{c}) > 0$ and $\frac{\partial df_p(c;\tilde{c})}{\partial c} > 0$ for $c \in (c_p, \tilde{c}]$.

Claim 3: (A.35) must be binding. This follows from a similar argument since $f_r > 0$ and

$f_r' > 0$.

Since (A.34) and (A.35) are binding, solving the integral equation in (A.34) with equality by taking the derivative gives (A.7) for $c \in (c_p, \tilde{c}]$. Plugging this solution into (A.35) and solving the integral equation by taking the derivative gives (A.7) for $c \in [\underline{c}, c_r)$.

Case 2: (A.5b) is binding. Substituting it for p_m in the objective function (A.5a) gives

$$\begin{aligned}\max_{W(\cdot), c_r \le c_p \le \tilde{c}} \quad & \Lambda F(\tilde{c}) V - \Lambda \int_{\underline{c}}^{c_r} W(c) \left[c + \frac{F(c)}{f(c)} \right] dF(c) - \Lambda \int_{c_p}^{\tilde{c}} W(c) \left[c + \frac{F(c)}{f(c)} \right] dF(c) \\ & - \Lambda \overline{W}(\tilde{c}) \left\{ c_p F(\tilde{c}) - \int_{c_p}^{\tilde{c}} \left[c - \frac{F(\tilde{c}) - F(c)}{f(c)} \right] dF(c) - \int_{\underline{c}}^{c_r} \left[c + \frac{F(c)}{f(c)} \right] dF(c) \right\} \end{aligned} \tag{A.36}$$

$$\text{s.t.} \qquad V - c_p \overline{W}(\tilde{c}) - \textstyle\int_{c_p}^{\tilde{c}} W(c) dc \le 0 \tag{A.37}$$

Now since for c in both intervals $[\underline{c}, c_r)$ and $(c_p, \tilde{c}]$, $W(c)$ is multiplied by the virtual type $f_r(c) = c + F(c)/f(c) > 0$, we claim similarly that the operational constraints (A.34) and (A.35) are binding. We show this for $c \in (c_p, \tilde{c}]$ and a similar argument applies to $c \in [\underline{c}, c_r)$. Suppose that (A.34) is not binding. Then there exists a feasible perturbation that decreases $W(c)$ for types $c \in (c_1, c_1 + \epsilon_1] \subset [c_p^*, \tilde{c}]$ and increases $W(c)$ for some types $c \in (c_2, c_2 + \epsilon_2] \subset [c_p^*, \tilde{c}]$ such that $\int_{c_p^*}^{\tilde{c}} W(c) dc$ is unchanged with $\epsilon_1, \epsilon_2 > 0$ and $c_2 + \epsilon_2 \le c_1$. This feasible perturbation strictly increases the objective function. Again, this holds since $f_r,\ f_r' > 0$. Following the claim, we can derive the expression of $W(c)$ as in Case 1.

Combining Case 1 and Case 2 gives part (i) of Lemma A.3.

Part (ii): Plugging the expressions in (A.7) into (A.35)
and recognizing that $\int_{\underline{c}}^{\tilde{c}} \frac{1}{\mu(1-\rho F(\tilde{c})+\rho F(c))^2} dF(c) = \frac{F(\tilde{c})}{\mu - \Lambda F(\tilde{c})}$ gives

$$Y \triangleq \int_{c_r}^{c_p} \frac{1}{\mu(1 - \rho F(\tilde{c}) + \rho F(c))^2} dF(c) - \frac{F(c_p) - F(c_r)}{\mu - \Lambda F(\tilde{c})} = 0. \tag{A.38}$$

Also,

$$\int_{c_r}^{c_p} \frac{1}{\mu(1-\rho F(\tilde{c})+\rho F(c))^2} dF(c) = \frac{F(c_p)-F(c_r)}{\mu(1-\rho F(\tilde{c})+\rho F(c_r))(1-\rho F(\tilde{c})+\rho F(c_p))} \tag{A.39}$$

Combining (A.38) and (A.39) yields (A.8).

Part (iii): Write $c_p(c_r)$ by way of substitution from (A.8). The optimization problem is stated with two decision variables p_m and c_r:

$$\begin{aligned} \max_{c_r, p_m} \quad \Pi(c_r, p_m) =& p_m \Lambda F(\tilde{c}) - \Lambda \int_{\underline{c}}^{c_r} (W^e(c;\tilde{c}) - \overline{W}(\tilde{c})) f_r(c) dF(c) \\ &+ \Lambda \int_{c_p(c_r)}^{\tilde{c}} (\overline{W}(\tilde{c}) - W^e(c;\tilde{c})) f_p(c;\tilde{c}) dF(c) \end{aligned} \tag{A.40}$$

$$\text{s.t.} \qquad V - c_p \overline{W}(\tilde{c}) - \int_{c_p}^{\tilde{c}} W^e(c;\tilde{c}) dc - p_m \geq 0, \tag{A.41}$$

$$- p_m \geq 0, \tag{A.42}$$

$$\underline{c} \leq c_r \leq c_p(c_r) \leq \tilde{c}. \tag{A.43}$$

Let $\gamma \geq 0$ be the dual multiplier for constraint (A.41) and $\pi \geq 0$ is the dual multiplier for constraint (A.42). The Lagrangian is

$$\begin{aligned} \mathcal{L} =& p_m \Lambda F(\tilde{c}) - \Lambda \int_{\underline{c}}^{c_r} (W^e(c;\tilde{c}) - \overline{W}(\tilde{c})) f_r(c) dF(c) + \Lambda \int_{c_p(c_r)}^{\tilde{c}} (\overline{W}(\tilde{c}) - W^e(c;\tilde{c})) f_p(c;\tilde{c}) dF(c) \\ &+ \gamma \left(V - c_p \overline{W}(\tilde{c}) - \int_{c_p}^{\tilde{c}} W^e(c;\tilde{c}) dc - p_m \right) - \pi p_m \end{aligned}$$

Taking the derivative of the Lagrangian with respect to c_r yields

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial c_r} =& -\Lambda (W^e(c_r;\tilde{c}) - \overline{W}(\tilde{c})) f_r(c_r) f(c_r) + \Lambda (W^e(c_p;\tilde{c}) - \overline{W}(\tilde{c})) f_p(c_p;\tilde{c}) f(c_p) \frac{dc_p}{dc_r} \\ &+ \gamma (-\overline{W}(\tilde{c}) + W^e(c_p;\tilde{c})) \frac{dc_p}{dc_r} \end{aligned} \tag{A.44}$$

From (A.38),

$$\frac{\partial Y}{\partial c_r} = -W^e(c_r;\tilde{c}) f(c_r) + f(c_r) \overline{W}(\tilde{c}), \qquad \frac{\partial Y}{\partial c_p} = W^e(c_p;\tilde{c}) f(c_p) - f(c_p) \overline{W}(\tilde{c}).$$

By the implicit function theorem,

$$\frac{dc_p}{dc_r} = -\frac{\frac{\partial Y}{\partial c_r}}{\frac{\partial Y}{\partial c_p}} = \frac{f(c_r)\left[W^e(c_r;\tilde{c}) - \overline{W}(\tilde{c})\right]}{f(c_p)\left[W^e(c_p;\tilde{c}) - \overline{W}(\tilde{c})\right]}. \tag{A.45}$$

Substituting A.45 into A.44 yields

$$\frac{\partial \mathcal{L}}{\partial c_r} = (W^e(c_r;\tilde{c}) - \overline{W}(\tilde{c}))f(c_r)\left[-\Lambda f_r(c_r) + \Lambda f_p(c_p;\tilde{c}) + \frac{\gamma}{f(c_p)}\right].$$

Thus setting $\frac{\partial \mathcal{L}}{\partial c_r} = 0$ gives

$$\Lambda f_r(c_r) = \Lambda f_p(c_p,\tilde{c}) + \frac{\gamma}{f(c_p)} \tag{A.46}$$

or

$$W^e(c_r;\tilde{c}) - \overline{W}(\tilde{c}) = 0 \tag{A.47}$$

Note that (A.47) corresponds to the case $c_r = c_p$. Since $\gamma \geq 0$, (A.46) gives $f_r(c_r^*(\tilde{c})) \geq f_p(c_p^*(\tilde{c}),\tilde{c})$. Note that this also holds if only (A.47) is true since (A.47) implies $c_r = c_p$.

In particular, if (A.5b) is not binding, then by complementary slackness, $\gamma = 0$, and therefore, (A.46) implies $f_r(c_r^*(\tilde{c})) = f_p(c_p^*(\tilde{c}),\tilde{c})$. We now verify $c_r = c_p$ implied by (A.47) cannot be optimal if (A.5b) is not binding. This is done by showing the second derivative is positive (locally convex) at the point where $c_r = c_p$. When $\gamma = 0$,

$$\frac{\partial^2 \mathcal{L}}{\partial c_r^2}|_{c_r:c_r=c_p} = \frac{dW^e(c_r;\tilde{c})}{dc_r} f(c_r)\left[-\frac{\Lambda F(\tilde{c})}{f(c_p)}\right] > 0$$

since $\frac{dW^e(c_r;\tilde{c})}{dc_r} < 0$.

Part (iv): Taking the partial derivative of the Lagrangian with respect to p_m gives

$$\frac{\partial \mathcal{L}}{\partial p_m} = \Lambda F(\tilde{c}) - \gamma - \pi = 0$$

Suppose $c_r < c_p$. Thus, (A.47) cannot hold and (A.46) holds. Since $f_r' > 0$,

$$\frac{f_r(c_r) - f_p(c_p;\tilde{c})}{F(\tilde{c})} < \frac{f_r(c_p) - f_p(c_p;\tilde{c})}{F(\tilde{c})} = \frac{1}{f_p(c_p;\tilde{c})}. \tag{A.48}$$

From (A.46), $\Lambda F(\tilde{c})\frac{f_r(c_r)-f_p(c_p;\tilde{c})}{F(\tilde{c})} = \frac{\gamma}{f_p(c_p;\tilde{c})}$. Combining this equation with (A.48) gives that $\Lambda F(\tilde{c}) > \gamma$. Therefore, $\pi > 0$. By complementary slackness, $p_m \leq 0$ must be binding: $p_m^* = 0$. □

Proof of Lemma A.4. This can be similarly proved as in part (i) of Lemma A.3 and is thus omitted. □

Proof of Lemma A.5. Refer to Problem (A.40)-(A.43) and now consider $\tilde{c}$ as a decision variable. Lemma A.3-(iv) shows that if $c_r < c_p$ then $p_m = 0$ for any $\tilde{c}$, so it must also hold for the optimal $\tilde{c}$. We now show that if $c_r^* = c_p^*$, it also always holds that $p_m^* = 0$. We first force $c_r = c_p$ and then show that the optimal solution to a problem without the constraint $p_m \leq 0$ in (A.42) will be infeasible to the problem with this constraint. Thus, this constraint must be binding.

If constraint (A.42) is removed, then the problem collapses to the service provider's problem (Problem A.3'). As we argue in the proof of Proposition 1.2, constraint (A.10b) must be binding. $P(\tilde{c}) = \tilde{p} = V - \frac{\tilde{c}}{\mu}$. By (A.2), the payment of the lowest type $\underline{c}$ is

$$\begin{aligned} P(\underline{c}) &= \tilde{p} - \tilde{c}W^e(\tilde{c};\tilde{c}) + \underline{c}W^e(\underline{c};\tilde{c}) - \int_{\underline{c}}^{\tilde{c}} W^e(c;\tilde{c})dc \\ &= V - \frac{\tilde{c}}{\mu} - \frac{2\rho}{\mu}\int_{\underline{c}}^{\tilde{c}} \frac{cf(c)}{[1-\rho F(\tilde{c})+\rho F(c)]^3}dc \quad \text{(integration by parts)}. \end{aligned}$$

Since the LHS of (A.20) (the derivative of the objective function in Problem A.3' w.r.t. $\tilde{c}$)

is nonnegative, by substitution,

$$\begin{aligned}
P(\underline{c}) =& V - \frac{\tilde{c}}{\mu} - \frac{2\rho}{\mu}\int_{\underline{c}}^{\tilde{c}} \frac{cf(c)}{[1-\rho F(\tilde{c})+\rho F(c)]^3}dc \\
=& \frac{F(\tilde{c})}{\mu f(\tilde{c})} + \frac{2\rho}{\mu}\int_{\underline{c}}^{\tilde{c}} \frac{cf(c)+F(c)}{[1-\rho F(\tilde{c})+\rho F(c)]^3}dc - \frac{2\rho}{\mu}\int_{\underline{c}}^{\tilde{c}} \frac{cf(c)}{[1-\rho F(\tilde{c})+\rho F(c)]^3}dc \\
=& \frac{F(\tilde{c})}{\mu f(\tilde{c})} + \frac{2\rho}{\mu}\int_{\underline{c}}^{\tilde{c}} \frac{F(c)}{[1-\rho F(\tilde{c})+\rho F(c)]^3}dc > 0,
\end{aligned}$$

Since $p_m \geq P(\underline{c})$, it follows that $p_m > 0$. This implies that constraint (A.42) is violated in the relaxed problem, which implies this constraint must be binding in the original problem. Hence, $p_m^* = 0$. □

A.6 Equilibrium of the H Auction When $H > \overline{H}$

In this section, we characterize the equilibrium when H is greater than $\overline{H}$ in the H auction. The proofs are in Appendix A.7.

We first derive the expected waiting time function if customers with type $c \in [c_r, c_p]$ are FIFO customers (like customer F in Illustration 1), and other customers are prioritized within themselves. This system is no longer a strict priority queue due to the presence of FIFO customers.

Lemma A.6. *If $J(c)$ =join for $c \in [\underline{c}, \tilde{c}]$ (and balk otherwise), customers in $[c_r, c_p]$ maintain their positions as if in a FIFO system, and the rest of joining customers are prioritized in decreasing order of their waiting cost c via trading, the expected waiting time for customer c is*

$$W^o(c; c_r, c_p, \tilde{c}) = \begin{cases} \frac{1-\rho(F(c_p)-F(c_r))}{\mu[1-\rho(F(\tilde{c})-F(c))]^2}, & \forall c \in [\underline{c}, c_r); \\ \overline{W}(\tilde{c}), & \forall c \in [c_r, c_p]; \\ \frac{1-\rho(F(c_p)-F(c_r))}{\mu[1-\rho(F(\tilde{c})-F(c))-\rho(F(c_p)-F(c_r))]^2}, & \forall c \in (c_p, \tilde{c}]. \end{cases} \tag{A.49}$$

The waiting time function of this system, $W^o(\cdot; c_r, c_p, \tilde{c})$ can be viewed as a generalization of the waiting time function of a priority queue, $W^e(\cdot; \tilde{c})$, given the same $\tilde{c}$. If $c_r = c_p$, then the two functions are identical. $W^o(\cdot; c_r, c_p, \tilde{c})$ is still strictly decreasing in $[\underline{c}, c_r) \cup (c_p, \tilde{c}]$ since these customers are prioritized among themselves. It is easy to check $W^o(c_r^-; c_r, c_p, \tilde{c}) = W^o(c_p^+; c_r, c_p, \tilde{c})$. This is because customers c_r^- and c_p^+ have adjacent priorities. For $c \in [\underline{c}, c_r)$, $W^o(c; c_r, c_p, \tilde{c}) < W^e(c; \tilde{c})$, implying that customers with low waiting cost in this system do not wait as long as they do in a priority queue. This result is driven by fewer overtaking from future arriving customers. On the other hand, customer $W^o(\tilde{c}; c_r, c_p, \tilde{c}) > W^e(\tilde{c}; \tilde{c}) = 1/\mu$, implying that the customer $\tilde{c}$ has to wait longer than in a priority queue. Originally, her expected waiting time is simply the service time; now she also has to wait behind some FIFO customers who refuse to give her priority. In other words, as compared to a strict priority queue, there is less disparity in the expected waiting time across customers in this system: those with low waiting costs do not wait as much, and those with high waiting cost do not wait as little. We use this waiting time function in finding the equilibrium of the auction when $H > \overline{H}$.

Theorem A.1 (Partial Trading Equilibrium)**.** *When $H > \overline{H}$, there exists an equilibrium in which:*

(i) $J^H(c)$ =join for $c \in [\underline{c}, \tilde{c}]$ (and balk otherwise);

(ii) the equilibrium bid function is given by

$$b^H(c; c_r, c_p, \tilde{c}) = \begin{cases} c + \frac{\int_c^{c_r} (F(\tilde{c}) - F(s) - F(c_p) + F(c_r))^2 W^o(s) ds + K_1}{(F(\tilde{c}) - F(c) - F(c_p) + F(c_r))^2 W^o(c)}, & \forall c \in [\underline{c}, c_r); \\ No, & \forall c \in [c_r, c_p]; \\ c + \frac{\int_c^{\tilde{c}} (F(\tilde{c}) - F(s))^2 W^o(s) ds}{(F(\tilde{c}) - F(c))^2 W^o(c)}, & \forall c \in (c_p, \tilde{c}]; \end{cases} \tag{A.50}$$

where $K_1(c_r, c_p, \tilde{c}) = (c_p - c_r)(F(\tilde{c}) - F(c_p))^2 W^o(c_r) + \int_{c_p}^{\overline{c}} (F(\tilde{c}) - F(s))^2 W^o(s) ds$ *and*

$c_r, c_p, \tilde{c} \in \Xi$ with $c_r < c_p$ are a solution to the following equations:

$$\overline{W}(\tilde{c}) - W^o(c_r^-; c_r, c_p, \tilde{c}) = 0 \tag{A.51}$$

$$\Pi^H(c_r, c_p, \tilde{c}) - H\Lambda[F(\tilde{c}) - F(c_p) + F(c_r)] = 0 \tag{A.52}$$

$$V - \int_{c_p}^{\tilde{c}} W^o(c; c_r, c_p, \tilde{c})dc - c_p\overline{W}(\tilde{c}) = 0 \tag{A.53}$$

where $\Pi^H(c_r, c_p, \tilde{c})$ denotes the intermediary's revenue and is given by $\Pi^H(c_r, c_p, \tilde{c}) =$

$$-\Lambda \int_{\underline{c}}^{c_r} (W^o(c; c_r, c_p, \tilde{c}) - \overline{W}(\tilde{c}))f_r(c)dF(c) + \Lambda \int_{c_p}^{\tilde{c}} (\overline{W}(\tilde{c}) - W^o(c; c_r, c_p, \tilde{c}))f_p(c; \tilde{c})dF(c),$$

and (A.53) is dropped if $\tilde{c} = \overline{c}$;
(iii) the expected waiting time of customer c is $W^o(c; c_r, c_p, \tilde{c})$.

Theorem A.1 shows two salient features of the equilibrium when the trade participation fee, H, is large:

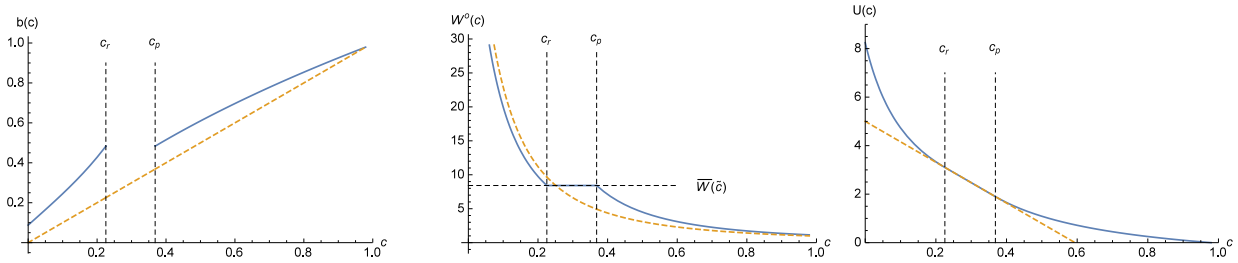
(1) Not all joining customers participate in trading;

(2) The schedule is not efficient.

These two features are in contrast to the equilibrium structure identified in Theorem 1.1 for the budget-balanced baseline auction. We explain these features via our numerical example (with parameters in Table 1.1) in Figure A.1 for $H = 1.446$. The customers in $[c_r, c_p]$ bid "No", maintain their waiting position as if in a FIFO system, and therefore expect the FIFO waiting time. Since customers with medium waiting costs are the ones who benefit the least from trading, thus if signing up for trading is too costly ($H > \overline{H}$), they would decide the benefit does not justify the cost and choose not to be involved in trading altogether. As illustrated in Figure A.1-(a), $b^H(c_r^-) = b^H(c_p^+)$, the two customers on the margin must bid equal price in equilibrium. Their equilibrium expected waiting time is also equal to the FIFO waiting time since $\overline{W}(\tilde{c}) = W^o(c_r^-)$ in (A.51). Note that this is a result

enforced by the equilibrium, not the waiting time function W^o itself. As indicated earlier, the scheduling policy as described in Lemma A.6 guarantees $W^o(c_r^-) = W^o(c_p^+)$ only, but does not imply W^o is continuous. However, in equilibrium W^o is continuous. It is clear from Figure A.1-(b) that the schedule is not efficient: W^o and W^e are different not only in $[c_r, c_p]$ but also outside of $[c_r, c_p]$. Even though trading does not change the non-trading customers' expected waiting time from the FIFO waiting time, trading customers are affected by the presence of non-trading customers as their expected waiting time now differs from what they would expect to get in an efficient schedule.

Figure A.1: An H auction with $H > \overline{H}$.



(a) Bidding function. (b) Waiting time. (c) Utility.

Note. $H = 1.446$. $\tilde{c} = 0.979$, $c_r = 0.225$, $c_p = 0.368$. The solid curve corresponds to the equilibrium properties of the auction; the dashed curve is the 45-degree line in (a); the efficient waiting time function in (b); the expected utility if customers are served FIFO in (c).

Note that in (A.52), customers in $[c_r, c_p]$ do not pay the trade participation fee H, so $H\Lambda[F(\tilde{c}) - F(c_p) + F(c_r)]$ is the intermediary's long-run average revenue. Thus $\Pi^H(c_r, c_p, \tilde{c})$ expresses the intermediary's revenue as a function of $c_r, c_p, \tilde{c}$. Equation (A.53) guarantees customers' join/balk decision is individually rational.

Figure A.1-(c) shows the utility curve in equilibrium. The utility of types in $[c_r, c_p]$ coincides with the utility obtained from FIFO joining because customers in $[c_r, c_p]$ do not participate in trading (but, do join). Figure A.1-(c) also gives intuitive interpretation of the conditions (A.51)-(A.53): customers on the margin, customers with waiting cost c_r and c_p must be both indifferent between trading and not trading and if some customers balk, the last customer who joins $\tilde{c}$ must be indifferent between joining and balking. Conditions

(A.51)-(A.53) are equivalent to:

$$U(c_r, b^H(c_r^-; c_r, c_p, \tilde{c})) = V - c_r \overline{W}(\tilde{c}) \tag{A.54}$$

$$U(c_p, b^H(c_p^+; c_r, c_p, \tilde{c})) = V - c_p \overline{W}(\tilde{c}) \tag{A.55}$$

$$U(\tilde{c}, b^H(\tilde{c}; c_r, c_p, \tilde{c})) = 0, \tag{A.56}$$

where $U(c, \beta)$ is the expected utility of customer c who bids β, including the trade participation fee if she chooses to pay.

In this equilibrium, any customer who bids lower than a new arriving customer will be pushed back by *more than* one spot in expectation while in all cases in the main body of Chapter 1, this number is equal to one. The rationale of this distinction is that here, we also need to account for the non-trading customers in $[c_r, c_p]$. Combining (A.51) and (A.54) implies that given the equilibrium bid function, any customer who bids $b^H(c_r^-)$ (or $b^H(c_p^+)$ since they are equal) makes *zero* expected net payment. The amount they receive when they sell should on average be equal to the amount they pay when they buy. By inspection, there always exists a solution to (A.51)-(A.53) with $c_r = \underline{c}$ and $c_p = \tilde{c}$. This is the trivial equilibrium that all joining customers bid "No." In fact, we find numerically that when H is sufficiently large, this is the only solution (which is not surprising). When H is greater than $\overline{H}$ by a moderate amount, equilibria with $\underline{c} < c_r < c_p < \tilde{c}$ arise. Continuing the numerical example in Figure A.1, three equilibria are found (see Table A.1), all of which involve at least some joining customers not participating in trading. In Equilibrium 1 the intermediary would earn 1.2614, the highest of the three. In Equilibrium 2 the intermediary would only earn 1.088 (14 % less); this equilibrium is also illustrated in Figure A.1. Equilibrium 3 is the trivial non-trading equilibrium. Note that in Equilibrium 1, the intermediary earns a higher revenue than if it charges $\overline{H} = 1.342$ (the corresponding revenue being 1.208); but this is still suboptimal as compared to the maximum revenue 1.352 under the optimal $(H, \underline{R}, \overline{R})$ auction (cf. Table 1.3).

Table A.1: Three different equilibria under $H = 1.446$.

	Equilibrium 1	Equilibrium 2	Equilibrium 3
Revenue	1.264	1.088	0
c_r	0.236	0.225	0
c_p	0.264	0.368	0.909
$\tilde{c}$	0.999	0.979	0.909

Note. Equilibrium 1 raises the maximum revenue for the intermediary among all H auctions.

A.7 Proofs of Lemma A.6 and Theorem A.1

Proof of Lemma A.6. We refer to customers in $[c_r, c_p]$ who do not participate in trading as FIFO customers and the rest of joining customers as the trading customers. The steady state system is organized as follows: head of the queue, a sequence of FIFO customers, a trading customer, a sequence of FIFO customers, a trading customer, etc. The trading customers are prioritized according to the $c\mu$ rule. Now we analyze the number of FIFO customers between two trading customers.

Claim A.1. *The number of FIFO customers in between two trading customers in the system follows an i.i.d. geometric distribution supported on* $\{0, 1, \ldots\}$ *with mean* $\frac{\rho[F(c_p)-F(c_r)]}{1-\rho[F(c_p)-F(c_r)]}$.

Proof of Claim A.1. The number of customers in the system follows a geometric distribution supported on $\{0, 1, \ldots\}$ and is equal to 0 with probability $1 - \tilde{\rho}$ where $\tilde{\rho} = \rho F(\tilde{c})$. Label each FIFO customer in the system as a red circle, and each trading customer as a blue circle. Then a circle is red with probability $p = \frac{F(c_p)-F(c_r)}{F(\tilde{c})}$ and blue with probability $1 - p$. This is equivalent to the following experiment: 1. Start with an empty strip; 2. With probability $1 - \tilde{\rho}$, the experiment ends; otherwise, draw a circle: fill it red with probability p, and blue with prob $(1 - p)$; 3. Keep iterating until the experiment ends. Therefore, the number of FIFO customers between two trading customers is equal to the number of consecutive red circles that following a blue circle. We keep drawing red circles until either the experiment ends or a blue circle is drawn. The union of these two events occur with probability $q = 1 - \tilde{\rho} + \tilde{\rho}(1 - p)$. The number of consecutive red circles is independent of whether the sequence ends because of a blue circle, or because the experiment ended. Thus

it follows a geometric distribution supported on $\{0, 1, \ldots\}$ with mean $\frac{1-q}{q}$. By substitution,

$$\frac{1-q}{q} = \frac{\tilde{\rho}p}{1-\tilde{\rho}p} = \frac{\rho[F(c_p) - F(c_r)]}{1-\rho[F(c_p) - F(c_r)]}. \qquad \square$$

Therefore, for customers $c \in [\underline{c}, c_r]$, the expected waiting time reduced and the expected waiting time increased are

$$\left(1 + \frac{\rho[F(c_p) - F(c_r)]}{1-\rho[F(c_p) - F(c_r)]}\right) \frac{\Lambda}{\mu} \int_{\underline{c}}^{c} W^o(s) dF(s) \text{ and}$$
$$\left(1 + \frac{\rho[F(c_p) - F(c_r)]}{1-\rho[F(c_p) - F(c_r)]}\right) \frac{\Lambda}{\mu} (F(\tilde{c}) - F(c_p) + F(c_r) - F(c)) W^o(c),$$

respectively. Combining the three contributions to the expected waiting time yields

$$\begin{aligned} W^o(c) =& \frac{1}{\mu-\lambda} - \left(1 + \frac{\rho[F(c_p) - F(c_r)]}{1-\rho[F(c_p) - F(c_r)]}\right) \frac{\Lambda}{\mu} \int_{\underline{c}}^{c} W^o(s) dF(s) \\ &+ \left(1 + \frac{\rho[F(c_p) - F(c_r)]}{1-\rho[F(c_p) - F(c_r)]}\right) \frac{\Lambda}{\mu} (F(\tilde{c}) - F(c_p) + F(c_r) - F(c)) W^o(c). \end{aligned}$$

Note that $\lambda = \Lambda F(\tilde{c})$. This simplifies to the following integral equation:

$$W^o(c) = \frac{1}{\mu-\lambda} - \int_{\underline{c}}^{c} \frac{\rho W^o(s)}{1-\rho(F(c_p) - F(c_r))} dF(s) + \frac{\rho W^o(c)}{1-\rho(F(c_p) - F(c_r))} (F(\tilde{c}) - F(c_p) + F(c_r) - F(c)).$$

Similar analysis can be done for $c \in (c_p, \tilde{c}]$ and thus the mean waiting time of type c customers is

$$W^o(c) = \begin{cases} \frac{1}{\mu-\lambda} - \int_{\underline{c}}^{c} \frac{\rho W^o(s)}{1-\rho(F(c_p)-F(c_r))} dF(s) & \\ + \frac{\rho W^o(c)}{1-\rho(F(c_p)-F(c_r))} (F(\tilde{c}) - F(c_p) + F(c_r) - F(c)), & \forall c \in [\underline{c}, c_r); \\ \frac{1}{\mu-\lambda}, & \forall c \in [c_r, c_p]; \\ \frac{1}{\mu-\lambda} - \int_{[\underline{c},c] \setminus [c_r,c_p]} \frac{\rho W^o(s)}{1-\rho(F(c_p)-F(c_r))} dF(s) + \frac{\rho W^o(c)}{1-\rho(F(c_p)-F(c_r))} (F(\tilde{c}) - F(c)), & \forall c \in (c_p, \tilde{c}]. \end{cases}$$

For $c \in [\underline{c}, c_r)$, $W(c)$ solves the following ODE:

$$W^{o\prime}(c) = -\frac{\rho W^o(c) f(c)}{1-\rho(F(c_p)-F(c_r))} + \frac{\rho W^{o\prime}(c)(F(\tilde{c})-F(c_p)+F(c_r)-F(c))}{1-\rho(F(c_p)-F(c_r))} - \frac{\rho W^o(c) f(c)}{1-\rho(F(c_p)-F(c_r))}$$

$$W^o(c) = K_1 \frac{1}{[1-\rho(F(\tilde{c})-F(c))]^2}.$$

The boundary condition at $\underline{c}$ gives

$$W^o(\underline{c}) = \frac{1}{\mu-\lambda} + \frac{\rho W^o(\underline{c})}{1-\rho(F(c_p)-F(c_r))}(F(\tilde{c})-F(c_p)+F(c_r)).$$

Solving this yields

$$W^o(\underline{c}) = \frac{1-\rho(F(c_p)-F(c_r))}{\mu(1-\rho F(\tilde{c}))^2}.$$

Therefore,

$$W^o(c) = \frac{1-\rho(F(c_p)-F(c_r))}{\mu[1-\rho(F(\tilde{c})-F(c))]^2}, \quad \forall c \in [\underline{c}, c_r)$$

For $c \in (c_p, \overline{c}]$, $W(c)$ solves the following ODE:

$$W^{o\prime}(c) = -\frac{\rho W^o(c) f(c)}{1-\rho(F(c_p)-F(c_r))} + \frac{\rho W^{o\prime}(c)(F(\tilde{c})-F(c))}{1-\rho(F(c_p)-F(c_r))} - \frac{\rho W^o(c) f(c)}{1-\rho(F(c_p)-F(c_r))}.$$

Solving this yields

$$W^o(c) = K_2 \frac{1}{[1-\rho(F(\tilde{c})-F(c))-\rho(F(c_p)-F(c_r))]^2}.$$

The boundary condition at $\tilde{c}$ gives

$$W^o(\tilde{c}) = \frac{1}{\mu} + \frac{\rho(F(c_p)-F(c_r))}{\mu[1-\rho(F(c_p)-F(c_r))]} = \frac{1}{\mu[1-\rho(F(c_p)-F(c_r))]}.$$

Therefore,

$$W^o(c) = \frac{1-\rho(F(c_p)-F(c_r))}{\mu[1-\rho(F(\tilde{c})-F(c))-\rho(F(c_p)-F(c_r))]^2}, \quad \forall c \in (c_p, \tilde{c}]. \qquad \square$$

Proof of Proposition A.1. The proof consists of four steps. We start with an equilibrium conjecture that there exists a cutoff delay cost $\tilde{c}$ such that only customers with type $c \in [\underline{c}, \tilde{c}]$ join the system (part (i)), and that for joining customers $c \in [\underline{c}, \tilde{c}]$, there exists a symmetric, symmetric bidding strategy $b^H(c; c_r, c_p, \tilde{c}),\ c \in [\underline{c}, \tilde{c}]$ that is increasing in $[\underline{c}, c_r)$ and $(c_p, \tilde{c}]$ respectively with $b^H(c_r^-; c_r, c_p, \tilde{c}) \leq b^H(c_p^+; c_r, c_p, \tilde{c})$, and "No" in $[c_r, c_p]$. We derive the equilibrium bid function (part (ii)) and waiting time (part (iii)) based on this conjecture in Step 1. We verify that this constitutes an equilibrium in Step 2-4. Specifically, we show (A.51)-(A.53) in Step 3.

Step 1: Equilibrium bid and waiting function. According to this equilibrium conjecture, the waiting time function will correspondingly be $W^o(c; c_r, c_p, \tilde{c})$ as derived in Lemma A.6.

We use the shorthand notation $b(\cdot)$ for the bid function and $W^o(\cdot)$ for the waiting time function whenever their meanings are clear. The expected utility of customer c who bids β is

$$U(c,\beta) = \begin{cases} V - cW^o(b^{-1}(\beta)) - \frac{1}{1-\rho(F(c_p)-F(c_r))} \int_{\underline{c}}^{b^{-1}(\beta)} \frac{b(s)}{\mu} \Lambda W^o(s) dF(s) & \\ +\frac{\beta}{\mu} \frac{1}{1-\rho(F(c_p)-F(c_r))} W^o(b^{-1}(\beta)) & \\ \cdot \Lambda[F(\tilde{c}) - F(b_H^{-1}(\beta)) - (F(c_p) - F(c_r))] - H, & \forall \beta \in [b(\underline{c}), b(c_r)); \\ V - cW^o(b^{-1}(\beta)) & \\ -\frac{1}{1-\rho(F(c_p)-F(c_r))} \int_{[\underline{c}, b^{-1}(\beta)] \setminus [c_r, c_p]} \frac{b(s)}{\mu} \Lambda W^o(s) dF(s) & \\ +\frac{\beta}{\mu} \frac{1}{1-\rho(F(c_p)-F(c_r))} W^o(b^{-1}(\beta)) \Lambda[F(\tilde{c}) - F(b^{-1}(\beta))] - H, & \forall \beta \in (b(c_p), b(\tilde{c})]; \\ V - c\overline{W}(\tilde{c}), & \beta = \text{No}. \end{cases}$$

The symmetric equilibrium strategy $b(c)$ satisfies the first-order condition $\frac{\partial U}{\partial \beta} \mid_{\beta = b(c)} = 0$ for $c \in [\underline{c}, c_r) \cup (c_p, \overline{c}]$ and the best response for customer c in $[c_r, c_p]$ should be to bid No.

For $c \in (c_p, \overline{c}]$,

$$\begin{aligned}
&\frac{\partial U}{\partial \beta} \mid_{\beta=b(c)} \\
=&\frac{-cW^{o\prime}(c)}{b'(c)} + \frac{-\rho b(c)W^o(c)f(c) + \rho b(c)W^o(c)(-f(c)) + \rho[W^o(c)b'(c) + b(c)W^{o\prime}(c)](1-F(c))}{b'(c)\left[1-\rho(F(c_p)-F(c_r))\right]} = 0.
\end{aligned}$$

After some algebra, this simplifies to the following linear differential equation:

$$[b'(c)-1](F(\tilde{c})-F(c))^2W^o(c) + [b(c)-c]\left\{(F(\tilde{c})-F(c))^2W^o(c)\right\}' = -(F(\tilde{c})-F(c))^2W^o(c). \tag{A.57}$$

The solution to (A.57) is given by

$$[b(c)-c](F(\tilde{c})-F(c))^2W^o(c) = \int_{\tilde{c}}^{c} -(F(\tilde{c})-F(s))^2W^o(s)ds + K,$$

where the constant K is chosen such that the right hand side is equal to zero at $c = \tilde{c}$; otherwise, $b(c)$ goes to infinity as c approaches $\tilde{c}$. This implies $K = 0$.

Similarly, for $c \in [\underline{c}, c_r)$, the first-order condition $\frac{\partial U}{\partial \beta} \mid_{\beta=b(c)} = 0$ simplifies to the following differential equation:

$$\begin{aligned}
&[b'(c)-1](F(\tilde{c})-F(c)-F(c_p)+F(c_r))^2W^o(c) \\
&+ [b(c)-c]\left\{(F(\tilde{c})-F(c)-F(c_p)+F(c_r))^2W^o(c)\right\}' \\
=&-(F(\tilde{c})-F(c)-F(c_p)+F(c_r))^2W^o(c).
\end{aligned} \tag{A.58}$$

The solution to (A.58) is given by

$$[b(c)-c](F(\tilde{c})-F(c)-F(c_p)+F(c_r))^2W^o(c) = \int_{c_r}^{c} -(F(\tilde{c})-F(s)-F(c_p)+F(c_r))^2W^o(s)ds + K_1,$$

We claim $b(c_r^-) = b(c_p^+)$. This gives us the boundary condition for pinning down K_1. We prove this claim by contradiction. Suppose $b(c_r^-) < b(c_p^+)$. Thus customer c_r would increase his expected utility by increasing his bid $b \in [b(c_r^-), b(c_p^+)]$. This is true because increasing

b within this range will not affect his expected waiting time, not affect the expected amount he pays for purchase, but strictly increases the amount he receives for selling spots. Thus $b(c_r^-) < b(c_p^+)$ is not supported in equilibrium. Therefore, $b(c_r^-) = b(c_p^+)$ gives the expression of K_1 as specified in Proposition A.1.

Step 2: Monotonicity. Since $b(c) > c$ for $c \in [\underline{c}, c_r)$ and $c \in (c_p, \overline{c}]$ respectively, differential equations (A.57) and (A.58) imply $b'(c) > 0$ for $c \in [\underline{c}, c_r)$ and $c \in (c_p, \overline{c}]$ respectively. Therefore, we show monotonicity of the bidding strategy in $[\underline{c}, c_r) \cup (c_p, \tilde{c}]$.

Step 3: Using IC and IR in joining to pin down $c_r, c_p, \tilde{c}$. It remains to verify that this is indeed an equilibrium by finding the correct $(c_r, c_p, \tilde{c})$: customers in $[\underline{c}, \tilde{c}]$ will not deviate from their own bid. This is shown in two steps: 1. no customer has an incentive to bid outside the range of $b(c)$: $[b(\underline{c}), b(\tilde{c})]$; 2. no customer has an incentive to send someone's else bid inside the range of $b(c)$ (including No), i.e., the auction is IC. The first step follows a similar argument in the proof of Theorem 1.1: bidding below $b(\underline{c})$ is less preferred than bidding $b(\underline{c})$ and there is no gain from bidding above $b(\tilde{c})$ as compared to bidding $b(\tilde{c})$. The second step is shown by the revenue equivalence theorem. By Lemma A.1, the auction is IC iff $W^o(c; \tilde{c})$ is weakly decreasing in c and

$$U(c) - U(c') = \int_c^{c'} W^o(s; c_r, c_p, \tilde{c}) ds. \tag{A.59}$$

Since $W^o(c; c_r, c_p, \tilde{c})$ is decreasing in $c \in [\underline{c}, c_r)$ and $(c_p, \tilde{c}]$ respectively and flat in $[c_r, c_p]$, $W^o(c; c_r, c_p, \tilde{c})$ will be weakly decreasing in c iff

$$W^o(c_r^-; c_r, c_p, \tilde{c}) \geq W(c_r; c_r, c_p, \tilde{c}), \quad \text{and} \quad W(c_p; c_r, c_p, \tilde{c}) \geq W^o(c_p^+; c_r, c_p, \tilde{c}).$$

Since $W^o(c_r^-; c_r, c_p, \tilde{c}) = W^o(c_p^+; c_r, c_p, \tilde{c})$ and $W^o(c_r; c_r, c_p, \tilde{c}) = \overline{W}(\tilde{c})$, $W^o(c_r^-; c_r, c_p, \tilde{c}) = \overline{W}(\tilde{c})$ in (A.51).

If $c_r = \underline{c}$, then $c_p = \tilde{c}$ by (A.51), and then (A.59) is trivially satisfied. This is an equilibrium where nobody trades. Otherwise, note that (A.59) is satisfied by the envelope

theorem for any $c, c' \in [\underline{c}, c_r) \cup (c_p, \tilde{c}]$. It is also satisfied by any $c, c' \in (c_r, c_p)$ since $U(c) - U(c') = (c' - c)\overline{W}(\tilde{c})$. We only need (A.59) to be satisfied at c_r and c_p, i.e., $U(c_r^-) = U(c_r)$ and $U(c_p^+) = U(c_p)$, where

$$U(c_r) = V - c_r \overline{W}(\tilde{c}),$$

$$\begin{aligned} U(c_r^-) =& V - c_r W^o(c_r^-; c_r, c_p, \tilde{c}) - \frac{\rho}{1 - \rho(F(c_p) - F(c_r))} \int_{\underline{c}}^{c_r^-} b(c) W^o(c; c_r, c_p, \tilde{c}) dF(c) \\ &+ \frac{\rho}{1 - \rho(F(c_p) - F(c_r))} b(c_r^-) W^o(c_r^-; c_r, c_p, \tilde{c})[F(\tilde{c}) - F(c_p)] - H. \end{aligned}$$

Since $W^o(c_r^-; c_r, c_p, \tilde{c}) = \overline{W}(\tilde{c})$, this implies

$$\begin{aligned} &\frac{\rho}{1 - \rho(F(c_p) - F(c_r))} \\ &\cdot \left\{ - \int_{\underline{c}}^{c_r^-} b(c) W^o(c; c_r, c_p, \tilde{c}) dF(c) + W^o(c_r^-; c_r, c_p, \tilde{c}) b(c_r^-)[F(\tilde{c}) - F(c_p)] \right\} - H = 0 \end{aligned} \tag{A.60}$$

Let $p^H(c)$ denote the net expected payment made by customer c excluding H for $c \in [\underline{c}, c_r) \cup (c_p, \tilde{c}]$. It is immediate that (A.60) can be rewritten as $p^H(c_r^-) = -H$. Equivalently, $p^H(c_p^+) = -H$. To simplify (A.60), note that the auction is budget-balanced excluding the payment to the intermediary H, i.e.,

$$\int_{\underline{c}}^{c_r} p^H(c) dF(c) + \int_{c_p}^{\tilde{c}} p^H(c) dF(c) = 0.$$

By (A.59) (guaranteed from the envelope theorem for $c \in [\underline{c}, c_r) \cup (c_p, \tilde{c}]$),

$$\int_{\underline{c}}^{c_r} p^H(c) dF(c) = p^H(c_r^-) F(c_r) + \int_{\underline{c}}^{c_r} \int_{c}^{c_r} s dW^o(s; c_r, c_p, \tilde{c}) dF(c),$$

$$\int_{c_p}^{\tilde{c}} p^H(c) dF(c) = p^H(c_p^+)[F(\tilde{c}) - F(c_p)] - \int_{c_p}^{\tilde{c}} \int_{c_p}^{c} s dW^o(s; c_r, c_p, \tilde{c}) dF(c).$$

Since $p^H(c_r^-) = p^H(c_p^+) = -H$, this implies

$$\begin{aligned} 0 =& \Lambda \int_{\underline{c}}^{c_r} p^H(c)dF(c) + \Lambda \int_{c_p}^{\tilde{c}} p^H(c)dF(c) \\ =& - H\Lambda[F(\tilde{c}) - F(c_p) + F(c_r)] + \Lambda \int_{\underline{c}}^{c_r} \int_{c}^{c_r} sdW^o(s; c_r, c_p, \tilde{c})dF(c) \\ & - \Lambda \int_{c_p}^{\tilde{c}} \int_{c_p}^{c} sdW^o(s; c_r, c_p, \tilde{c})dF(c). \end{aligned} \tag{A.61}$$

One can show by similar techniques in deriving the intermediary's revenue in (A.5a) that

$$\Lambda\Pi^H(c_r, c_p, \tilde{c}) = -\Lambda \int_{\underline{c}}^{c_r} \int_{c}^{c_r} sdW^o(s; c_r, c_p, \tilde{c})dF(c) + \Lambda \int_{c_p}^{\tilde{c}} \int_{c_p}^{c} sdW^o(s; c_r, c_p, \tilde{c})dF(c). \tag{A.62}$$

Combining (A.61) and (A.62) gives (A.52). Note that (A.52) subsumes that case where $c_r = \underline{c}$ and $c_p = \tilde{c}$.

The threshold joining strategy follows from the same argument as in Theorem 1.1. Furthermore, if $\tilde{c} < \bar{c}$, then IR in joining requires (A.56). This can be simplified to (A.53) in a similar way as we derive (A.9c).

By inspection, there exists at least one solution to (A.51)-(A.53), which has $c_r = \underline{c}$ and $c_p = \tilde{c}$. Also, any solution cannot have $c_r = c_p$ since this would imply $H = \overline{H}$ from (A.52), which contradicts the condition $H > \overline{H}$.

Step 4: IR in trading. Since customers in $[c_r, c_p]$ do not trade, $U(c) = V - c\overline{W}(\tilde{c})$ for $c \in [c_r, c_p]$. Since we use IC to pin down c_r, c_p, $U(c)$ is continuous and $U'(c) = -W^o(c)$ almost everywhere. Since $W^o(c) > \overline{W}(\tilde{c})$ for $c \in [\underline{c}, c_r)$ and $W^o(c) < \overline{W}(\tilde{c})$ for $c \in (c_p, \tilde{c}]$, we have $U(c) > V - c\overline{W}(\tilde{c})$ for $c \in [\underline{c}, c_r) \cup (c_p, \tilde{c}]$. Therefore, $U(c) \geq V - c\overline{W}(\tilde{c}),\ \forall c \in [\underline{c}, \tilde{c}]$. $\square$

APPENDIX B

SUPPLEMENT TO CHAPTER 2

B.1 Finding Equilibria and Proof of Theorem 2.1

In this appendix, we develop a constructive method that allows computing all pure-strategy equilibria, and mixed strategy equilibria if pure-strategy equilibria do not exist.

Pure Strategy Equilibria. For any conjecture of the reservation utility $\bar{U} = u$ in the relevant range, recall from (2.5) that any equilibrium search thresholds $\boldsymbol{k}$ must satisfy:

$$u \in \left[V_i - \frac{c}{\mu}(k_i+1), V_i - \frac{c}{\mu}k_i\right], \text{ if } k_i > 0; \quad u \geq V_i - \frac{c}{\mu}, \text{ if } k_i = 0, \quad i = 1, \ldots, N.$$

Thus, with a slight abuse of notation, we define the mapping $\boldsymbol{k}(\cdot) : \mathbb{R}_+ \mapsto \mathbb{N}^N$ such that $\boldsymbol{k}(u) = (k_1(u), ..., k_N(u))$, where

$$k_i(u) = \left(\left\lfloor (V_i - u)\frac{\mu}{c} \right\rfloor\right)^+,$$

with $(x)^+ = \max\{0, x\}$. For quality level i, $k_i(u)$ defined above is rational under conjecture $\bar{U} = u$.

Consider $\bar{U}(\boldsymbol{k}, \boldsymbol{\pi}(\boldsymbol{k}))$ from Equation (2.4). Define the mapping $F : \mathbb{R}_+ \mapsto \mathbb{R}_+$ such that $F(u) \triangleq \bar{U}(\boldsymbol{k}(u), \boldsymbol{\pi}(\boldsymbol{k}(u)))$. Any pure strategy equilibrium is characterized by a fixed point of mapping F. If u^* satisfies

$$u^* = F(u^*), \tag{B.1}$$

then the equilibrium search thresholds are $\boldsymbol{k}(u^*)$. Consequently, we translate the N-dimensional fixed point problem to a single-dimensional one.

Next we provide more details about $F(u)$. First, $F(u)$ is a step function since each of $k_i(u)$ is a step function. Define $u_j^i = V_i - \frac{c}{\mu}(j+1)$ for $i \in \{1, ..., N\}$ and $j \in \{0, ..., \lfloor \frac{V_i \mu}{c} \rfloor\}$. At

$u = u_j^i$, $k_i(u)$ decreases by one: $k_i(u_j^i+) = k_i(u_j^i-) - 1$. By construction, $k_i(0) = \bar{k}_i = \lfloor \frac{V_i\mu}{c} \rfloor$ (defined in Assumption 2.1). Therefore, Assumption 2.1 implies $F(0) > 0$. We next quantify the domain of u in which $F(u)$ is well posed, i.e., Equation (2.2) admits a solution $\alpha(\boldsymbol{k}(u))$. Lemma 2.1 requires $\sum_{i\in\mathcal{S}} p_i > \rho$, where $\mathcal{S} = \{i|k_i > 0\}$. Since k_i is nondecreasing in quality level i, we obtain the following characterization. Let i^* be the index such that $\sum_{i=i^*}^{N} p_i > \rho$, $\sum_{i=i^*+1}^{N} p_i \leq \rho$. Denote $\bar{u} = V_{i^*} - c/\mu$. Thus, the domain of $F(u)$ is $u \in [0, \bar{u}]$.

We sort the collection $\{u_j^i\}_{i\in\{1,...,N\},\ j\in\{0,...,\lfloor \frac{V_i\mu}{c} \rfloor\}}$ from lowest to greatest. Denote the sorted set as $\mathcal{U} \triangleq \{u_1, u_2, ..., u_L\}$ (in which duplicates are removed). Note that $u_L = \bar{u}$. At every u_l, $l \in \{1, ..., L\}$, there is at least one component of $\boldsymbol{k}(u)$ that decreases by 1 at $u = u_l$. $F(u)$ is a step function: let $F(u) = F_{l+1}$ for $u \in (u_l, u_{l+1}]$, $l = 0...L-1$, where we let $u_0 = 0$.

Mixed Strategy Equilibria. If Equation (B.1) does not have a fixed point, then there is no pure strategy equilibrium, and we look for mixed strategy equilibria. There are two cases. Case (i): $F(\bar{u}) < \bar{u}$. Then since $F(0) > 0$, there must exist l^* such that either $F_{l^*} < u_{l^*} < F_{l^*+1}$ or $F_{l^*} > u_{l^*} > F_{l^*+1}$. Define an index set $\mathcal{I} = \{i|(V_i - u_{l^*})\frac{\mu}{c} \in \mathbb{N}\}$. Case (ii): $F(\bar{u}) > \bar{u}$. We examine these two cases and establish the existence of mixed strategy equilibria in either case in Proposition B.1. This shows the existence of either pure or mixed strategy equilibria, and hence we prove Theorem 2.1 by proving Proposition B.1.

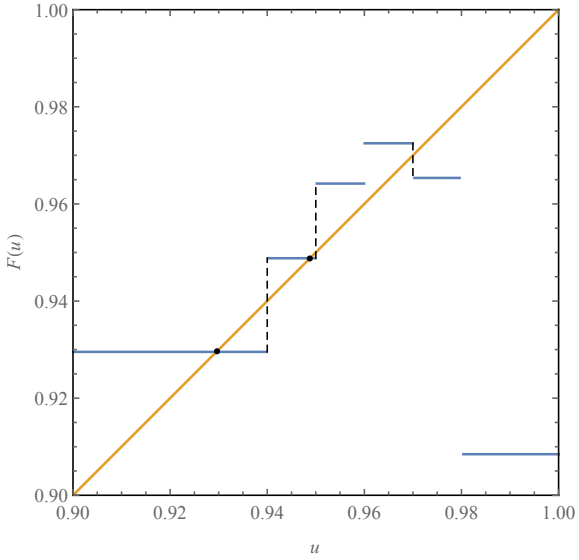
Proposition B.1. *A mixed-strategy equilibrium exists in Cases (i) and (ii):*

- *Case (i) There exists a mixed strategy equilibrium in which $k_i^* = k_i(u_{l^*})$ for $i \notin \mathcal{I}$ and $k_i^* = k_i(u_{l^*}) - 1 + \kappa^*$ for $i \in \mathcal{I}$ with $\kappa^* \in (0, 1)$.*
- *Case (ii) There exists a mixed strategy equilibrium in which $k_i^* = k_i(\bar{u})$ for $i \neq i^*$ and $k_{i^*}^* = \kappa^*$ with $\kappa^* \in (0, 1)$.*

Example B.1. *Consider the following parameters: $\mu = 1$; $\lambda = 0.801$; $s = 0.3155$; $c = 0.26$ and $N = 5$. The quality levels are given by $V_i = \frac{3}{2} + \frac{i-1}{N-1}$ and $p_i = \frac{1}{5}$ for all $i \in \{1, ..., 5\}$.*

There are two pure-strategy equilibria: (P1) is $(2,3,4,5,6)$ *and (P2) is* $(2,3,4,5,5)$*. There are three mixed-strategy equilibria: (M1) is* $(2,3,4,5,5.402)$*, (M2) is* $(2,3,4,4.902,5)$*, and (M3) is* $(2,2.546,3,4,5)$*.* $(M3)$ *is the Pareto-dominant equilibrium.*

Figure B.1: Finding equilibria for model primitives given in Example B.1.



Note. There are five equilibria: two pure-strategy equilibria, designated by filled circles satisfying $F(u) = u$, and three mixed-strategy equilibria, designated by the intersections of $F(u)$ and the vertical dashed lines.

Proof of Proposition B.1. For notational convenience, we suppress the dependence of k_i on u_l or $\bar{u}$ whenever the meanings are clear.

Case (i): We prove the case $F_{l^*} < u_{l^*} < F_{l^*+1}$. The other case $F_{l^*} > u_{l^*} > F_{l^*+1}$ is analogous.

Given the search thresholds $(\boldsymbol{k}(u_{l^*}), \kappa)$, the queue length distribution is given:

$$\pi^i_j(\boldsymbol{k},\kappa,\alpha) = \frac{\alpha^j}{\sum_{l=0}^{k_i}\alpha^l}, i \notin \mathcal{I},$$

$$\pi^{\tilde{\imath}}_j(\boldsymbol{k},\kappa,\alpha) = \frac{\alpha^j}{\sum_{l=0}^{k_{\tilde{\imath}}-1}\alpha^l + \kappa\alpha^{k_{\tilde{\imath}}}},\ j < k_{\tilde{\imath}} \text{ and } \pi^{\tilde{\imath}}_{k_{\tilde{\imath}}}(\boldsymbol{k},\kappa,\alpha) = \frac{\kappa\alpha^{k_{\tilde{\imath}}}}{\sum_{l=0}^{k_{\tilde{\imath}}-1}\alpha^l + \kappa\alpha^{k_{\tilde{\imath}}}},\ \tilde{\imath} \in \mathcal{I}.$$

Here, α must also satisfy:

$$\frac{\alpha}{\rho} = \frac{1}{1 - \sum_{\tilde{\imath}\in\mathcal{I}} p_{\tilde{\imath}} \frac{(1-\kappa)\alpha^{k_{\tilde{\imath}}-1} + \kappa\alpha^{k_{\tilde{\imath}}}}{\sum_{l=0}^{k_{\tilde{\imath}}-1}\alpha^l + \kappa\alpha^{k_{\tilde{\imath}}}} - \sum_{i\notin\mathcal{I}} p_i \frac{\alpha^{k_i}}{\sum_{l=0}^{k_i}\alpha^l}}. \tag{B.2}$$

The reservation utility is

$$\begin{aligned}\bar{U}(\boldsymbol{k},\kappa,\alpha) &= \frac{-s + \sum_{i\notin\mathcal{I}} p_i \sum_{j=0}^{k_i-1} \pi_j^i(\boldsymbol{k},\kappa,\alpha)(V_i - c\frac{j+1}{\mu})}{1 - \sum_{\tilde{\imath}\in\mathcal{I}} p_{\tilde{\imath}}[\pi^{\tilde{\imath}}_{k_{\tilde{\imath}}}(\boldsymbol{k},\kappa,\alpha) + (1-\kappa)\pi^{\tilde{\imath}}_{k_{\tilde{\imath}}-1}(\boldsymbol{k},\kappa,\alpha)] - \sum_{i\notin\mathcal{I}} p_i \pi^i_{k_i}(\boldsymbol{k},\kappa,\alpha)} \\ &+ \frac{\sum_{\tilde{\imath}\in\mathcal{I}} p_{\tilde{\imath}}[\sum_{j=0}^{k_{\tilde{\imath}}-2} \pi_j^{\tilde{\imath}}(\boldsymbol{k},\kappa,\alpha)(V_{\tilde{\imath}} - c\frac{j+1}{\mu}) + \kappa\pi^{\tilde{\imath}}_{k_{\tilde{\imath}}-1}(\boldsymbol{k},\kappa,\alpha)(V_{\tilde{\imath}} - c\frac{k_{\tilde{\imath}}}{\mu})]}{1 - \sum_{\tilde{\imath}\in\mathcal{I}} p_{\tilde{\imath}}[\pi^{\tilde{\imath}}_{k_{\tilde{\imath}}}(\boldsymbol{k},\kappa,\alpha) + (1-\kappa)\pi^{\tilde{\imath}}_{k_{\tilde{\imath}}-1}(\boldsymbol{k},\kappa,\alpha)] - \sum_{i\notin\mathcal{I}} p_i \pi^i_{k_i}(\boldsymbol{k},\kappa,\alpha)}.\end{aligned}$$

Randomizing at queue length $k_{\tilde{\imath}} - 1$ for quality level $\tilde{\imath} \in \mathcal{I}$ is rational only when customers are indifferent between joining or searching at queue length $k_{\tilde{\imath}} - 1$ for servers of quality $\tilde{\imath}$. Therefore,

$$\bar{U}(\boldsymbol{k},\kappa,\alpha) = V_{\tilde{\imath}} - \frac{c}{\mu}k_{\tilde{\imath}}. \tag{B.3}$$

We need to prove that there exist $\alpha^* > 0$ and $\kappa^* \in (0,1)$ that simultaneously solve Equations (B.2) and (B.3).

Let $\boldsymbol{k}^L$ be $\boldsymbol{k}(u_{l^*} - \varepsilon)$ and $\boldsymbol{k}^R$ be $\boldsymbol{k}(u_{l^*} + \varepsilon)$ where $\varepsilon > 0$ is a small positive number. Recall that $k_i^L = k_i^R$ for all $i \notin \mathcal{I}$ and $k_{\tilde{\imath}}^R = k_{\tilde{\imath}}^L - 1$ for $i \in \mathcal{I}$. Denote

$$U^m(\alpha^m) = \frac{-s + \sum_{i=1}^N p_i \sum_{j=0}^{k_i^m-1} \pi_j^i(\boldsymbol{k}^m,\alpha^m)(V_i - c\frac{j+1}{\mu})}{1 - \sum_{i=1}^N p_i \pi^i_{k_i^m}(\boldsymbol{k}^m,\alpha^m)}, \quad m \in \{L,R\},$$

where $\pi_j^i(\boldsymbol{k}^m,\alpha^m) = \frac{\alpha^j}{\sum_{l=0}^{k_i}\alpha^l}$, and α^m is the positive root α of $\frac{\alpha}{\rho} = \frac{1}{1 - \sum_{i=1}^N p_i \frac{\alpha^{k_i^m}}{\sum_{l=0}^{k_i^m}\alpha^l}}$, $m \in \{L,R\}$.

Note that $F_{l^*} = U^L(\alpha^L)$ and $F_{l^*+1} = U^R(\alpha^R)$. Since $F_{l^*} < u_{l^*+1} < F_{l^*+1}$, $U^L(\alpha^L) < V_{\tilde{\imath}} - \frac{c}{\mu}k_{\tilde{\imath}} = \bar{U}(\boldsymbol{k},\kappa,\alpha) < U^R(\alpha^R)$.

Given $\boldsymbol{k} = \boldsymbol{k}^L$, write the solution α in Equation (B.2) as a function of κ, $\hat{\alpha} : [0,1] \mapsto \mathbb{R}^+$.

It is immediate that $\hat{\alpha}(\kappa)$ is continuous. By inspection, $\hat{\alpha}(0) = \alpha^R$, and $\hat{\alpha}(1) = \alpha^L$. By some algebra, $\bar{U}(\boldsymbol{k}^L, 0, \hat{\alpha}(0)) = U^R(\alpha^R)$, and $\bar{U}(\boldsymbol{k}^L, 1, \hat{\alpha}(1)) = U^L(\alpha^L)$. Since $\bar{U}(\boldsymbol{k}^L, \kappa, \hat{\alpha}(\kappa))$ is continuous in κ and $U^L(\alpha^L) < V_{\tilde{\imath}} - \frac{c}{\mu}k_{\tilde{\imath}} < U^R(\alpha^R)$, therefore, by the intermediate value theorem, there must exist a $\kappa^* \in (0,1)$ such that $\hat{U}(\boldsymbol{k}^L, \kappa^*, \hat{\alpha}(\kappa^*)) = V_{\tilde{\imath}} - \frac{c}{\mu}k_{\tilde{\imath}}$, satisfying Equation (B.3). Let $\alpha^* = \hat{\alpha}(\kappa^*)$. Then, (κ^*, α^*) simultaneously solve Equations (B.2) and (B.3).

Case (ii): Let $\boldsymbol{k} = \boldsymbol{k}(\bar{u})$. Thus,

$$\bar{U}(\boldsymbol{k}, \kappa, \alpha) = \frac{\alpha}{\rho}\left\{-s + \sum_{i=i^*+1}^{N} p_i \sum_{j=0}^{k_i-1} \pi_j^i(\boldsymbol{k}, \kappa, \alpha)\left(V_i - c\frac{j+1}{\mu}\right) + p_{i^*}\frac{\kappa}{1+\kappa\alpha}\left(V_{i^*} - \frac{c}{\mu}\right)\right\},$$

where α solves the equation:

$$\rho = \sum_{i=i^*+1}^{N} p_i \left(1 - \frac{1}{\sum_{j=0}^{k_i}\alpha^j}\right) + p_{i^*}\frac{\kappa\alpha}{1+\kappa\alpha}. \tag{B.4}$$

Let $\hat{\alpha}(\kappa)$ be the solution α to Equation (B.4) for a given $\kappa > 0$. First recall that if $\kappa = 0$, there is no admissible solution to Equation (B.4) by definition of i^*. As $\kappa \to 0^+$, $\hat{\alpha}(\kappa)$ tends to infinity, which implies $\lim_{\kappa\to 0^+}\sum_{i=i^*+1}^{N} p_i\left(1 - \frac{1}{\sum_{j=0}^{k_i}\hat{\alpha}(\kappa)^j}\right) = \sum_{i=i^*+1}^{N} p_i$. Hence, by Equation (B.4),

$$\lim_{\kappa\to 0^+} p_{i^*}\frac{\kappa\hat{\alpha}(\kappa)}{1+\kappa\hat{\alpha}(\kappa)} = \rho - \sum_{i=i^*+1}^{N} p_i. \tag{B.5}$$

This further implies that

$$\lim_{\kappa\to 0^+}\frac{\hat{\alpha}(\kappa)}{\rho}\left\{\sum_{i=i^*+1}^{N}p_i\sum_{j=0}^{k_i-1}\pi_j^i(\boldsymbol{k},\kappa,\hat{\alpha}(\kappa))\left(V_i-c\frac{j+1}{\mu}\right)+p_{i^*}\frac{\kappa}{1+\kappa\hat{\alpha}(\kappa)}\left(V_{i^*}-\frac{c}{\mu}\right)\right\}$$

$$=\lim_{\kappa\to 0^+}\frac{1}{\rho}\sum_{i=i^*+1}^{N}p_i\sum_{j=0}^{k_i-1}\frac{\hat{\alpha}(\kappa)^{j+1}}{\sum_{j=0}^{k_i}\hat{\alpha}(\kappa)^j}\left(V_i-c\frac{j+1}{\mu}\right)+\frac{1}{\rho}p_{i^*}\frac{\kappa\hat{\alpha}(\kappa)}{1+\kappa\hat{\alpha}(\kappa)}\left(V_{i^*}-\frac{c}{\mu}\right)$$

$$=\frac{1}{\rho}\sum_{i=i^*+1}^{N}p_i\left(V_i-c\frac{k_i}{\mu}\right)+\left(1-\frac{1}{\rho}\sum_{i=i^*+1}^{N}p_i\right)\left(V_{i^*}-\frac{c}{\mu}\right)\qquad\text{By (B.5) and }\hat{\alpha}(0^+)\to\infty$$

$$=\left(V_{i^*}-\frac{c}{\mu}\right)+\frac{1}{\rho}\sum_{i=i^*+1}^{N}p_i\left[\left(V_i-c\frac{k_i}{\mu}\right)-\left(V_{i^*}-\frac{c}{\mu}\right)\right]\qquad\text{Since }V_{i^*}-\frac{c}{\mu}=\bar{u}\text{ and }V_i-c\frac{k_i}{\mu}<\bar{u}+\frac{c}{\mu}$$

$$<\bar{u}+\frac{c}{\lambda}\sum_{i=i^*+1}^{N}p_i.$$

Since $s>0$ and $\hat{\alpha}(0^+)\to\infty$, $\bar{U}(\boldsymbol{k},\kappa,\hat{\alpha}(\kappa))\to-\infty$ as $\kappa\to 0^+$. Since $\bar{U}(\boldsymbol{k},1,\hat{\alpha}(1))=F(\bar{u})>\bar{u}$. By the intermediate value theorem, there exists $\kappa\in(0,1)$ for which $\bar{U}(\boldsymbol{k},\kappa,\hat{\alpha}(\kappa))=\bar{u}=V_{i^*}-c/\mu$. □

B.2 Other Proofs

Proof of Lemma 2.1. Consider the expected change in the number of servers of quality i with j customers over a small period of time of length dt. The probability a customer arrives during this period is $n\lambda dt$. Two events due to arrivals can change $n_j^i(t,\boldsymbol{k})$ for $0\le j\le k_i$, $i\in\{1,\ldots,N\}$. One (for $j<k_i$) is that an arriving customer joins a server of quality i with j customers, which occurs with probability $\frac{n_j^i(t,\boldsymbol{k})}{n-\sum_{i=1}^{N}n_{k_i}^i(t,\boldsymbol{k})}$; this decreases $n_j^i(t,\boldsymbol{k})$ by 1. The other (for $j>0$) is that an arriving customer joins a server of quality i with $j-1$ customers, which increases $n_j^i(t,\boldsymbol{k})$ by 1. Thus, the expected change in $n_j^i(t,\boldsymbol{k})$ for $0<j<k_i$ due to arrivals is $-n\lambda\frac{n_j^i(t,\boldsymbol{k})-n_{j-1}^i(t,\boldsymbol{k})}{n-\sum_{i=1}^{N}n_{k_i}^i(t,\boldsymbol{k})}dt$. Two events due to departures can change $n_j^i(t,\boldsymbol{k})$. One (for $j<k_i$) is that a customer leaves a server of quality i with j customers, which occurs with probability $\mu n_j^i(t,\boldsymbol{k})dt$ for $j>0$; this leads to a decrease in $n_j^i(t,\boldsymbol{k})$ by 1. The other (for $j>0$) is that the departures of a customer at servers of quality i with $j+1$ customers,

which occurs with probability $\mu n^i_{j+1}(t,\boldsymbol{k})$; this increases $n^i_j(t,\boldsymbol{k})$ by 1. Thus, the expected change in $n^i_j(t,\boldsymbol{k})$ for $0 < j < k_i$ due to departures is $\mu(n^i_{j+1}(t,\boldsymbol{k}) - n^i_j(t,\boldsymbol{k}))dt$. The cases where $j = 0$ and $j = k_i$ should be handled with slightly extra care since no customers would leave an empty server and no customers would join a "full" queue.

Hence, when $\sum_{i=1}^N n^i_{k_i}(t,\boldsymbol{k}) < n$, the time evolution of the infinite system behaves according to these expectations and is determined by the following set of ordinary differential equations (ODEs):

$$\sum_{i=1}^N n^i_{k_i}(t,\boldsymbol{k}) < n : \begin{bmatrix} \dot{n}^i_0(t,\boldsymbol{k}) \\ \vdots \\ \dot{n}^i_j(t,\boldsymbol{k}) \\ \vdots \\ \dot{n}^i_{k_i}(t,\boldsymbol{k}) \end{bmatrix} = \begin{bmatrix} \mu n^i_1(t,\boldsymbol{k}) - n\lambda \frac{n^i_0(t,\boldsymbol{k})}{n - \sum_{i=1}^N n^i_{k_i}(t,\boldsymbol{k})} \\ \vdots \\ \mu(n^i_{j+1}(t,\boldsymbol{k}) - n^i_j(t,\boldsymbol{k})) - n\lambda \frac{n^i_j(t,\boldsymbol{k}) - n^i_{j-1}(t,\boldsymbol{k})}{n - \sum_{i=1}^N n^i_{k_i}(t,\boldsymbol{k})} \\ \vdots \\ -\mu n^i_{k_i}(t,\boldsymbol{k}) + n\lambda \frac{n^i_{k_i-1}(t,\boldsymbol{k})}{n - \sum_{i=1}^N n^i_{k_i}(t,\boldsymbol{k})} \end{bmatrix}, \ \forall\ 1 \le i \le N; \tag{B.6}$$

If $\sum_{i=1}^N N^i_{k_i}(\boldsymbol{k}) = n$, which implies an arriving customer, after exhaustive search, does not wish to join any server, the system states can only change due to service completion. This leads to system (B.7).

$$\sum_{i=1}^N n^i_{k_i}(t,\boldsymbol{k}) = n : \begin{bmatrix} \dot{n}^i_0(t,\boldsymbol{k}) \\ \vdots \\ \dot{n}^i_{k_i-2}(t,\boldsymbol{k}) \\ \dot{n}^i_{k_i-1}(t,\boldsymbol{k}) \\ \dot{n}^i_{k_i}(t,\boldsymbol{k}) \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \mu n^i_{k_i}(t,\boldsymbol{k}) \\ -\mu n^i_{k_i}(t,\boldsymbol{k}) \end{bmatrix}, \ \forall\ 1 \le i \le N. \tag{B.7}$$

For each i, dividing both sides by n_i gives the ODEs that characterize the time eveolution

of $\boldsymbol{\pi}(t,\boldsymbol{k})$:

$$\sum_{i=1}^{N}\pi^i_{k_i}(t,\boldsymbol{k})<1:$$

$$\begin{bmatrix}\dot\pi^i_0(t,\boldsymbol{k})\\ \vdots\\ \dot\pi^i_j(t,\boldsymbol{k})\\ \vdots\\ \dot\pi^i_{k_i}(t,\boldsymbol{k})\end{bmatrix}=\begin{bmatrix}\pi^i_1(t,\boldsymbol{k})\mu-\frac{\pi^i_0(t,\boldsymbol{k})\lambda}{1-\sum_{i=1}^N p_i\pi^i_{k_i}(t,\boldsymbol{k})}\\ \vdots\\ (\pi^i_{j+1}(t,\boldsymbol{k})-\pi^i_j(t,\boldsymbol{k}))\mu-\frac{(\pi^i_j(t,\boldsymbol{k})-\pi^i_{j-1}(t,\boldsymbol{k}))\lambda}{1-\sum_{i=1}^N p_i\pi^i_{k_i}(t,\boldsymbol{k})}\\ \vdots\\ -\pi^i_{k_i}(t,\boldsymbol{k})\mu+\frac{\pi^i_{k_i-1}(t,\boldsymbol{k}))\lambda}{1-\sum_{i=1}^N p_i\pi^i_{k_i}(t,\boldsymbol{k})}\end{bmatrix},\ \forall\ 1\le i\le N; \tag{B.8}$$

$$\sum_{i=1}^N\pi^i_{k_i}(t,\boldsymbol{k})=1:\begin{bmatrix}\dot n^i_0(t,\boldsymbol{k})\\ \vdots\\ \dot\pi^i_{k_i-2}(t,\boldsymbol{k})\\ \dot\pi^i_{k_i-1}(t,\boldsymbol{k})\\ \dot\pi^i_{k_i}(t,\boldsymbol{k})\end{bmatrix}=\begin{bmatrix}0\\ \vdots\\ 0\\ \mu\\ -\mu\end{bmatrix},\ \forall\ 1\le i\le N. \tag{B.9}$$

To find $\boldsymbol{\pi}(\boldsymbol{k})$, the fixed point of $\boldsymbol{\pi}(t,\boldsymbol{k})$, let $\dot\pi^i_j(t,\boldsymbol{k})=0$ for $1\le j\le k_i$, $1\le i\le N$. It is immediate from (B.9) that $\sum_{i=1}^N\pi^i_{k_i}(t,\boldsymbol{k})=1$ cannot be a fixed point. From (B.8), $\pi^i_1(\boldsymbol{k})=\alpha\pi^i_0(\boldsymbol{k})$ and $(\pi^i_{j+1}(\boldsymbol{k})-\pi^i_j(\boldsymbol{k}))=\alpha(\pi^i_j(\boldsymbol{k})-\pi^i_{j-1}(\boldsymbol{k}))$ for $1\le j\le k_i-1$ and $\pi^i_{k_i}(\boldsymbol{k})=\alpha\pi^i_{k_i-1}(\boldsymbol{k})$, where $\alpha=\frac{\rho}{1-\sum_{i=1}^N p_i\pi^i_{k_i}(\boldsymbol{k})}$. This gives $\pi^i_j(\boldsymbol{k})=\alpha^j\pi^i_0(\boldsymbol{k})$. Since $\sum_{j=0}^{k_i}\pi^i_j(\boldsymbol{k})=1$, $\pi^i_0(\boldsymbol{k})=1/\sum_{j=0}^{k_i}\alpha^j$. Thus,

$$\pi^i_j(\boldsymbol{k})=\frac{\alpha^j}{\sum_{l=0}^{k_i}\alpha^l}\qquad\text{for } 0\leqslant j\leqslant k_i, i\in\{1,\dots,N\}.$$

This gives (2.1). Substituting this into the defining equation of α yields the condition for α:

$$\alpha(\boldsymbol{k}):\alpha=\frac{\rho}{1-\sum_{i=1}^N p_i\frac{\alpha^{k_i}}{\sum_{l=0}^{k_i}\alpha^l}}.$$

This gives (2.2). Since $\pi_j^i(\boldsymbol{k})$ are between 0 and 1, α is positive. Therefore it is the positive root of (2.2). Next we show the existence and uniqueness of the positive root. Let $\mathcal{N} = \{1, 2, \ldots, N\}$. Thus shuffling terms of (2.2) yields

$$f(\alpha) \triangleq \sum_{i \in \mathcal{N} \setminus \mathcal{S}} p_i + \sum_{i \in \mathcal{S}} \frac{p_i}{\sum_{l=0}^{k_i} \alpha^l} - (1 - \rho) = 0.$$

For $\alpha > 0$, $f(\alpha)$ declines with α. $f(0) = +\infty$ and $f(\infty) = \rho - \sum_{i \in \mathcal{S}} p_i < 0$. Hence, there exists a unique positive solution α. □

Next, we introduce definitions, notation, some useful properties of stochastic ordering and likelihood ratio ordering (see, e.g., Shaked and Shanthikumar (2007), pp. 4 and pp. 42). These concepts will be used in later proofs.

Definition B.1 (Stochastic Ordering)**.** *If* $\mathbb{P}(X \leq a) \geq \mathbb{P}(Y \leq a)$, $\forall a$, *then* $X \leq_{st} Y$.

Definition B.2 (Likelihood Ratio Ordering)**.** *If* X, Y *are discrete random variables, and* $\frac{\mathbb{P}(Y=a)}{\mathbb{P}(X=a)}$ *is nondecreasing in* a, *then* $X \leq_{lr} Y$.

Property 1: If $X \leq_{lr} Y$, then $X \leq_{st} Y$.

Property 2: If $X \leq_{st} Y$, then $\mathbb{E}[X] \leq \mathbb{E}[Y]$.

Lemma B.1 characterizes some important properties of the special case of identical quality $N = 1$. Let the random variable $Q(k)$ be the queue length observed after a uniformly random sample of servers if customers all follow the search threshold k. Thus, $\mathbb{P}(Q(k) = j) = \pi_j^1(k) \triangleq \pi_j(k)$. We drop the superscript for quality in $\pi_j(k)$ since $N = 1$. Let $EQ(k)$ be the average queue length in the system: $EQ(k) = \sum_{j=0}^{k} j\pi_j(k)$.

Lemma B.1. *For* $N = 1$, $\forall k \geq 1$,

(i) $\alpha(k) > \alpha(k+1)$;

(ii) $Q(k) \leq_{st} Q(k+1)$;

(iii) $EW(k) < EW(k+1)$;

(iv) $\pi_k(k) > \pi_{k+1}(k+1)$;

(v) $\pi_{k+1}(k+1) < EQ(k+1) - EQ(k) < \rho$.

Proof. **Proof of Lemma B.1** *Part (i) and (ii)*: From (2.2), $\pi_0(k) = 1-\rho$, $\forall k$, and $\alpha(k) > \alpha(k+1)$. From (2.1),

$$\pi_j(k) = \pi_0(k)[\alpha(k)]^j = (1-\rho)[\alpha(k)]^j,$$

$\pi_j(k) > \pi_j(k+1), \forall 1 \le j \le k$. Thus, $\sum_{l=0}^{j} \pi_j(k) \ge \sum_{l=0}^{j} \pi_j(k+1)$ $\forall 0 \le j \le k$, where the equality holds iff $j = 0$. Therefore, $Q(k) \le_{st} Q(k+1)$.

Part (iii): It follows from (ii) that $EQ(k) < EQ(k+1)$. Thus, by Little's Law, $EW(k) < EW(k+1)$.

Part (iv): Since $\alpha(k) = \frac{\rho}{1-\pi_k(k)}$, $\alpha(k) > \alpha(k+1)$ gives $\pi_k(k) > \pi_{k+1}(k+1)$.

Part (v): Note that $EQ(k) = \sum_{j=0}^{k}(1 - F_j(k))$, where $F_j(k) = \sum_{l=0}^{j} \pi_j(k)$ is the CDF. Since $F_k(k) = 1$, $EQ(k) = \sum_{j=0}^{k-1}(1 - F_j(k))$. $EQ(k+1) - EQ(k) = \sum_{j=0}^{k}(1 - F_j(k+1)) - \sum_{j=0}^{k-1}(1 - F_j(k))$. Note that since $\pi_0(k) = 1-\rho$ $\forall k \ge 1$, $1 - F_0(k) = 1 - F_0(k+1) = \rho$. Hence,

$$EQ(k+1) - EQ(k) = \sum_{j=1}^{k-1}(F_j(k) - F_j(k+1)) + \pi_{k+1}(k+1).$$

Since $\sum_{j=1}^{k-1}(F_j(k) - F_j(k+1)) > 0$ from part (i), $EQ(k+1) - EQ(k) > \pi_{k+1}(k+1)$.

$$EQ(k+1) - \rho = \sum_{j=0}^{k+1} j\pi_j(k+1) - \rho = \sum_{j=0}^{k} j\pi_{j+1}(k+1) < \frac{\sum_{j=0}^{k} j\pi_{j+1}(k+1)}{1 - \pi_0(k+1)}.$$

The second equality follows from $\sum_{j=1}^{k+1} \pi_j(k+1) = \rho$. From (2.1) and by some algebra,

$$\frac{\sum_{j=0}^{k} j\pi_{j+1}(k+1)}{1 - \pi_0(k+1)} = \frac{\sum_{j=0}^{k} j[\alpha(k+1)]^j}{\sum_{j=0}^{k}[\alpha(k+1)]^j}.$$

Note that

$$EQ(k) = \frac{\sum_{j=0}^{k} j[\alpha(k)]^j}{\sum_{j=0}^{k}[\alpha(k)]^j} > \frac{\sum_{j=0}^{k} j[\alpha(k+1)]^j}{\sum_{j=0}^{k}[\alpha(k+1)]^j}.$$

The inequality holds since $\alpha(k) > \alpha(k+1)$ and $\frac{\sum_{j=0}^{k} jx^j}{\sum_{j=0}^{k} x^j}$ is increasing in x for $x > 0$.

Consequently, $\pi_{k+1}(k+1) < EQ(k+1) - EQ(k) < \rho, \forall k \geq 1$. □

Proof of Theorem 2.2. The proof is carried out in two steps. In the first step, we show that pure strategy equilibria always exist; in the second step, we show that a mixed strategy equilibrium cannot be Pareto dominant.

Step 1: Since there is only one quality level, the dynamic program for sequential search (2.3) can be simplified to a minimization problem (given other customers' search threshold k):

$$C(k) = \frac{s\mu}{c} + \sum_{j=0}^{k} \pi_j(k) \min\{j+1, C(k)\} \tag{B.10}$$

where $C(k)$ is the normalized "reservation cost," i.e., $C(k) = \frac{\mu}{c}(V_1 - \bar{U})$. A pure strategy equilibrium k^* must satisfy $k^* \leq C(k^*) \leq k^* + 1$.

Observe from (B.10) that $C(k) \leq 1 + \frac{s\mu}{c} + EQ(k)$. Since $1 + \frac{s\mu}{c} + EQ(1) = 1 + \frac{s\mu}{c} + \rho > 1$ and by Lemma B.1-(v), $0 < (1 + \frac{s\mu}{c} + EQ(k+1)) - (1 + \frac{s\mu}{c} + EQ(k)) < \rho < 1$, there exists a k such that $k \leq 1 + \frac{s\mu}{c} + EQ(k) < k + 1$. Let $\underline{k}^*$ be the smallest k that satisfies this, i.e., $\underline{k}^* \leq 1 + \frac{s\mu}{c} + EQ(\underline{k}^*) < \underline{k}^* + 1$ and $1 + \frac{s\mu}{c} + EQ(\underline{k}^* - 1) \geq \underline{k}^*$. We shall show $\underline{k}^* \leq C(\underline{k}^*) < \underline{k}^* + 1$, which implies that $\underline{k}^*$ is the search threshold, and hence the existence of a pure-strategy equilibrium.

We first consider the case $\underline{k}^* = 1$, i.e., $1 \leq 1 + \frac{s\mu}{c} + EQ(1) < 2$. This implies $\frac{s\mu}{c} + \rho < 1$. Suppose that $C(1) \in [1, 2)$. From (B.10), $C(1) = 1 + \frac{s\mu}{c(1-\rho)}$. Since $\frac{s\mu}{c} + \rho < 1$, $1 + \frac{s\mu}{c(1-\rho)} \in [1, 2)$, and this verifies $C(1) \in [1, 2)$. Thus, $\underline{k}^* = 1$ is a pure-strategy equilibrium.

Next, we consider $\underline{k}^* > 1$. If $\underline{k}^*$ is a pure-strategy equilibrium, i.e., $\underline{k}^* \leq C(\underline{k}^*) < \underline{k}^* + 1$, then by (B.10) $C(\underline{k}^*) = 1 + \frac{s\mu}{c} + EQ(\underline{k}^*) - (\underline{k}^* + 1)\pi_{\underline{k}^*}(\underline{k}^*) + \pi_{\underline{k}^*}(\underline{k}^*)C(\underline{k}^*) \in [\underline{k}^*, \underline{k}^* + 1)$.

Let $\psi(x) \triangleq 1 + \frac{s\mu}{c} + EQ(\underline{k}^*) - (\underline{k}^* + 1)\pi_{\underline{k}^*}(\underline{k}^*) + \pi_{\underline{k}^*}(\underline{k}^*)x - x$. An equilibrium k^* should satisfy $\psi(C(\underline{k}^*)) = 0$ and $C(\underline{k}^*) \in [\underline{k}^*, \underline{k}^* + 1)$. Now we prove such k^* can be found.

Since $1 + \frac{s\mu}{c} + EQ(\underline{k}^*) < \underline{k}^* + 1$, $\psi(\underline{k}^* + 1) < 0$. On the other hand, $\psi(\underline{k}^*) = 1 + \frac{s\mu}{c} + EQ(\underline{k}^*) - (\underline{k}^* + 1)\pi_{\underline{k}^*}(\underline{k}^*) + \pi_{\underline{k}^*}(\underline{k}^*)\underline{k}^* - \underline{k}^* = 1 + \frac{s\mu}{c} + EQ(\underline{k}^*) - \pi_{\underline{k}^*}(\underline{k}^*) - \underline{k}^*$. By Lemma B.1-(v), $EQ(\underline{k}^*) - \pi_{\underline{k}^*}(\underline{k}^*) > EQ(\underline{k}^* - 1)$, and since $1 + \frac{s\mu}{c} + EQ(\underline{k}^* - 1) \geq \underline{k}^*$; therefore, $\psi(\underline{k}^*) > 0$. Hence, by the intermediate value theorem, there exists $C(\underline{k}^*) \in (\underline{k}^*, \underline{k}^* + 1)$ that solves $\psi(C(\underline{k}^*)) = 0$. Hence, $\underline{k}^*$ is the equilibrium search threshold. This completes the proof of existence of pure-strategy equilibria.

Step 2: If a mixed strategy equilibrium is a Pareto dominant equilibrium, then according to the transformation in Appendix B.1, there must exists u, such that $F(u) > u$, and for all $u' > u$, $F(u') < u'$ (like the Pareto dominant equilibrium in Example B.1. See Figure B.1). Since there exists at least one pure strategy equilibrium u^* that solves $F(u^*) = u^*$, it must be that $u^* < u$. We show that this is impossible by showing that for any $u' < u$, $F(u') > u'$.

Let $\gamma(k) \triangleq \frac{\alpha(k)s\mu}{c\rho} + \frac{EQ(k)}{\rho}$. Thus, the equilibrium k satisfies $k \leq \gamma(k) < k+1$. Note that $\gamma(k(u)) = \mu(V - F(u))/c$, where $k(u) = \lfloor \mu(V-u)/c \rfloor$. Thus, showing that if $F(u) > u$, then for any $u' < u$, $F(u') > u'$ is equivalent to showing that if $\gamma(\tilde{k}) < \tilde{k}$, then $\gamma(k) < k$, $\forall\ k \geq \tilde{k}$. Note that Since $\alpha(k)$ is decreasing in k, $\gamma(k+1) - \gamma(k) < \frac{EQ(k+1) - EQ(k)}{\rho}$. From Lemma B.1-(v), $EQ(k+1) - EQ(k) < \rho$, we have $\gamma(k+1) - \gamma(k) < 1$. Therefore, if $\gamma(\tilde{k}) < \tilde{k}$, then $\gamma(k) < k$, $\forall\ k \geq \tilde{k}$. This completes the proof.

□

We define several system metrics to be used in later proofs.

Let $\phi_i(\boldsymbol{k})$ denote the fraction of customers who join quality i servers given search thresholds $\boldsymbol{k}$. For pure strategy $\boldsymbol{k}$,

$$\phi_i(\boldsymbol{k}) = \frac{p_i(1 - \pi^i_{k_i}(\boldsymbol{k}))}{1 - \sum_{q=1}^{N} p_q \pi^q_{k_q}(\boldsymbol{k})}.$$

Note that the average quality customers obtain given $\boldsymbol{k}$ can also be written by

$$EV(\boldsymbol{k}) = \sum_{i=1}^{N} \phi_i(\boldsymbol{k}) V_i.$$

Let $EW_i(\boldsymbol{k})$ denote the average waiting time conditional on joining a server of quality i given search thresholds $\boldsymbol{k}$. For pure strategy $\boldsymbol{k}$,

$$EW_i(\boldsymbol{k}) \triangleq \frac{\sum_{j=0}^{k_i-1} \pi_j^i(\boldsymbol{k})(j+1)/\mu}{1-\pi_{k_i}^i(\boldsymbol{k})}.$$

Note that the unconditional average waiting time in the system can also be written by

$$EW(\boldsymbol{k}) \triangleq \sum_{i=1}^{N} \phi_i(\boldsymbol{k}) EW_i(\boldsymbol{k}).$$

Proof of Proposition 2.1. Let random variable $V(\boldsymbol{k})$ denotes the quality a customer obtains in equilibrium $\boldsymbol{k}$ with $\mathbb{P}(V(\boldsymbol{k}) = V_i) = \phi_i(\boldsymbol{k}),\ i = 1, \ldots, N$. Let V be the random variable following the original quality distribution with $\mathbb{P}(V = V_i) = p_i,\ i = 1, \ldots, N$. We shall show $V \leq_{lr} V(\boldsymbol{k})$.

For any quality i with a pure-strategy search threshold,

$$\begin{aligned}\frac{\phi_i(\boldsymbol{k})}{p_i} &= \frac{\alpha(\boldsymbol{k})}{\rho}\left[1 - \frac{\alpha(\boldsymbol{k})^{k_i}}{\sum_{j=0}^{k_i}\alpha(\boldsymbol{k})^j}\right] = \frac{1}{\rho}\left[\alpha(\boldsymbol{k}) - \frac{\alpha(\boldsymbol{k})^{k_i+1}}{\sum_{j=0}^{k_i}\alpha(\boldsymbol{k})^j}\right] \\ &= \frac{1}{\rho}\left[\frac{\sum_{j=1}^{k_i}\alpha(\boldsymbol{k})^j + \alpha(\boldsymbol{k})^{k_i+1}}{\sum_{j=0}^{k_i}\alpha(\boldsymbol{k})^j} - \frac{\alpha(\boldsymbol{k})^{k_i+1}}{\sum_{j=0}^{k_i}\alpha(\boldsymbol{k})^j}\right] = \frac{1}{\rho}\left[\frac{\sum_{j=1}^{k_i}\alpha(\boldsymbol{k})^j}{\sum_{j=0}^{k_i}\alpha(\boldsymbol{k})^j}\right] = \frac{1}{\rho}\left[1 - \frac{1}{\sum_{j=0}^{k_i}\alpha(\boldsymbol{k})^j}\right].\end{aligned}$$

Since $1 - \frac{1}{\sum_{j=0}^{k_i}\alpha(\boldsymbol{k})^j}$ is increasing in k_i for fixed $\alpha(\boldsymbol{k})$ and k_i is nondecreasing in i, it follows that $\frac{\phi_i(\boldsymbol{k})}{p_i}$ is nondecreasing in quality level i with a pure-strategy search threshold.

For any quality i with a mixed-strategy search threshold,

$$\frac{\phi_i(\boldsymbol{k})}{p_i} = \frac{\alpha(\boldsymbol{k})}{\rho}\left[1 - \frac{(1-\kappa)\alpha(\boldsymbol{k})^{k_i-1} + \kappa\alpha(\boldsymbol{k})^{k_i}}{\sum_{j=0}^{k_i-1}\alpha(\boldsymbol{k})^j + \kappa\alpha(\boldsymbol{k})^{k_i}}\right] \in \frac{\alpha(\boldsymbol{k})}{\rho}\left[1 - \frac{\alpha(\boldsymbol{k})^{k_i-1}}{\sum_{j=0}^{k_i-1}\alpha(\boldsymbol{k})^j}, 1 - \frac{\alpha(\boldsymbol{k})^{k_i}}{\sum_{j=0}^{k_i}\alpha(\boldsymbol{k})^j}\right].$$

Hence, we can apply the same logic as in the pure-strategy case and conclude that $\frac{\phi_i(\boldsymbol{k})}{p_i}$ is nondecreasing in quality level i for both pure-strategy and mixed-strategy equilibria.

This implies $V \leq_{lr} V(\boldsymbol{k})$. It immediately follows that $EV^* \geq \sum_{i=1}^{N} p_i V_i$. □

Proof of Proposition 2.2. The proof is carried out in four steps. Step 1 shows part (ii) and specifies the conditions under which this constitutes a Pareto dominant equilibrium. Step 2 proves part (i), (iii), (iv). Step 3 proves part (v). Step 4 proves that $\overline{s}$ is increasing in integer d_N. We suppress the superscript $*$ to denote equilibrium for notational convenience.

Step 1: We focus on the equilibrium structure in which customers adopt search thresholds $\boldsymbol{k} = (\kappa, k_2, \ldots, k_N)$, $\kappa \in (0,1)$ (joining a server of the lowest quality level with probability κ if it currently has no customers and search otherwise; joining a server of quality i if its queue length less than k_i and search otherwise). This is an equilibrium if the following conditions are satisfied:

$$\frac{\rho}{\alpha(\boldsymbol{k})} = 1 - p_1 \frac{\kappa\alpha(\boldsymbol{k}) + 1 - \kappa}{1 + \kappa\alpha(\boldsymbol{k})} - \sum_{i=2}^{N} p_i \frac{\alpha(\boldsymbol{k})^{k_i}}{\sum_{j=0}^{k_i} \alpha(\boldsymbol{k})^j}, \tag{B.11}$$

$$V_1 - \frac{c}{\mu} = \frac{\alpha(\boldsymbol{k})}{\rho} \left\{ -s + p_1 \frac{\kappa}{1 + \kappa\alpha(\boldsymbol{k})} \left(V_1 - \frac{c}{\mu} \right) + \sum_{i=2}^{N} p_i \frac{\sum_{j=0}^{k_i - 1} \alpha(\boldsymbol{k})^j \left(V_i - c\frac{j+1}{\mu} \right)}{\sum_{l=0}^{k_i} \alpha(\boldsymbol{k})^l} \right\} = \bar{U}(\boldsymbol{k}), \tag{B.12}$$

$$V_i - c\frac{k_i + 1}{\mu} \leq V_1 - \frac{c}{\mu} \leq V_i - c\frac{k_i}{\mu}, \quad i = 2, \ldots, N. \tag{B.13}$$

Condition (B.12) specifies that customers are indifferent between joining an empty low server and the reservation utility from search in a mixed-strategy equilibrium. Condition (B.13) implies $k_i = 1 + \lfloor d_i \rfloor$ for $i = 2, \ldots, N$. It is also immediate that $\bar{U}(\boldsymbol{k}) \equiv V_1 - c/\mu$ (part (ii)) as s changes (as long as this equilibrium structure $(\kappa, k_2, \ldots, k_N)$ is maintained). Conditions

(B.11)-(B.13) simplify to

$$\rho = p_1 \frac{\kappa\alpha(\boldsymbol{k})}{1+\kappa\alpha(\boldsymbol{k})} + \sum_{i=2}^{N} p_i \left(1 - \frac{1}{\sum_{j=0}^{1+\lfloor d_i \rfloor} \alpha(\boldsymbol{k})^j}\right), \tag{B.14}$$

$$s = \frac{c}{\mu} \sum_{i=2}^{N} p_i \frac{\sum_{j=0}^{\lfloor d_i \rfloor} \alpha(\boldsymbol{k})^j \, (d_i - j)}{\sum_{l=0}^{1+\lfloor d_i \rfloor} \alpha(\boldsymbol{k})^l} \triangleq \frac{c}{\mu} \eta(\alpha(\boldsymbol{k})). \tag{B.15}$$

To satisfy conditions (B.14) and (B.15), there must exist a solution $(\kappa, \alpha(\boldsymbol{k}))$ to this system of simultaneous equations. We shall show that if $(1-p_1)\mu < \lambda$, then for any $s < \bar{s}$, such a solution exists. We shall explicitly specify $\bar{s}$. we start by showing $\eta(\alpha)$ is decreasing in $\alpha \in (0, \infty)$.

$$\eta(\alpha) = \sum_{i=2}^{N} p_i \left(\sum_{j=0}^{\lfloor d_i \rfloor - 1} \frac{\sum_{l=0}^{j} \alpha^l}{\sum_{l=0}^{1+\lfloor d_i \rfloor} \alpha^l} + (d_i - \lfloor d_i \rfloor) \frac{\sum_{l=0}^{\lfloor d_i \rfloor} \alpha^l}{\sum_{l=0}^{1+\lfloor d_i \rfloor} \alpha^l} \right). \tag{B.16}$$

It suffices to show that for positive d, $\frac{\sum_{l=0}^{j} \alpha^l}{\sum_{l=0}^{1+\lfloor d \rfloor} \alpha^l}$ is decreasing in α for any $j = 0, 1, \ldots, \lfloor d \rfloor$. It further suffices to show $\frac{\sum_{l=0}^{1+\lfloor d \rfloor} \alpha^l}{\sum_{l=0}^{j} \alpha^l}$ is increasing in α. Since

$$\frac{\sum_{l=0}^{1+\lfloor d \rfloor} \alpha^l}{\sum_{l=0}^{j} \alpha^l} = 1 + \frac{\sum_{l=j+1}^{1+\lfloor d \rfloor} \alpha^l}{\sum_{l=0}^{j} \alpha^l},$$

It suffices to show $\frac{\alpha^{j'}}{\sum_{l=0}^{j} \alpha^l}$ is increasing in α for $j' > j$. Diving both the numerator and the denominator by $\alpha^{j'}$ gives $\frac{1}{\sum_{l=0}^{j} \alpha^{l-j'}}$, which is increasing in α. This proves $\eta(\alpha)$ is decreasing in α. From (B.15), this implies that as search cost s decreases, $\alpha(\boldsymbol{k})$ increases. Furthermore, from (B.14), increasing $\alpha(\boldsymbol{k})$ decreases κ. Therefore, κ increases in s: Hence, $\bar{s}$ is determined by setting $\kappa = 1$:

$$\bar{s} = \frac{c}{\mu}\eta(\bar{\alpha}) \quad \text{where } \bar{\alpha} \text{ solves } \rho = \sum_{i=1}^{N} p_i \left(1 - \frac{1}{\sum_{j=0}^{1+\lfloor d_i \rfloor} \bar{\alpha} j}\right). \tag{B.17}$$

For any $s < \bar{s}$, one can solve for $\alpha(\boldsymbol{k})$ from (B.15), plug it into (B.14) to solve for κ. Since $(1-p_1) < \rho$, any κ found in this way will satisfy $\kappa > 0$. Thus, there always exist a valid solution $(\kappa, \alpha(\boldsymbol{k}))$ to (B.14) and (B.15) for $s < \bar{s}$. The equilibrium is sustained. Moreover, this is a Pareto dominant equilibrium. Suppose that there is an equilibrium with $\bar{U} > V_1 - c/\mu$. It must be that customers do not join the lowest quality severs at all. Hence $k_1 = 0$. However, since $(1-p_1) < \rho$, the system will not be stable, which leads to contradiction.

Step 2: Since $\alpha(\boldsymbol{k})$ increases as s decreases, $ES(\boldsymbol{k}) = \alpha(\boldsymbol{k})/\rho$ also increases, which shows part (i). For any $i \geq 2$, $\phi_i(\boldsymbol{k}) = p_i \left(1 - \frac{1}{\sum_{j=0}^{1+\lfloor d_i \rfloor} \alpha(\boldsymbol{k})^j}\right)/\rho$, which is increasing in $\alpha(\boldsymbol{k})$. Since $\alpha(\boldsymbol{k})$ increases as s decreases, therefore, $\phi_i(\boldsymbol{k})$ increases for any $i \geq 2$. Hence, $EV = \sum_{i=1}^{N} \phi_i(\boldsymbol{k}) V_i$ increases (part (iii)). Since $EW_i(\boldsymbol{k})$ is increasing in $\alpha(\boldsymbol{k})$ for $i \geq 2$ and $EW_1(\boldsymbol{k}) = 1/\mu$, $EW(\boldsymbol{k}) = \sum_{i=1}^{N} \phi_i(\boldsymbol{k}) EW_i(\boldsymbol{k})$ is increasing in $\alpha(\boldsymbol{k})$ and thus increases as s decreases (part (iv)).

Step 3: Since $ER(\boldsymbol{k}) = \bar{U}(\boldsymbol{k}) + sES(\boldsymbol{k})$ and $\bar{U}(\boldsymbol{k})$ is unchanged in s, part(v) to equivalent to showing $sES(\boldsymbol{k})$ decreases as s decreases.

$$sES(\boldsymbol{k}) = s\frac{\alpha(\boldsymbol{k})}{\rho} \quad = \quad \frac{c}{\mu\rho}\alpha(\boldsymbol{k})\eta(\alpha(\boldsymbol{k})).$$

Since $\alpha(\boldsymbol{k})$ increases as s decreases, we need to provide conditions under which $\alpha\eta(\alpha)$ is decreasing in α to show that $sES(\boldsymbol{k})$ decreases as search cost falls. From (B.16), for integer d,

$$\alpha\eta(\alpha) = \sum_{i=2}^{N} p_i \sum_{j=0}^{d_i - 1} \frac{\sum_{l=0}^{j} \alpha^{l+1}}{\sum_{l=0}^{1+d_i} \alpha^l} = \sum_{i=2}^{N} p_i \sum_{j=1}^{d_i} \frac{\sum_{l=1}^{j} \alpha^l}{\sum_{l=0}^{1+d_i} \alpha^l}.$$

By Lemma B3 in Hu et al. (2016), for any $j = 1, 2, \ldots, d$, $\frac{\sum_{l=1}^{j} \alpha^l}{\sum_{l=0}^{1+d} \alpha^l}$ is strictly decreasing in α for $\alpha \geq 1$. Hence, $\alpha\eta(\alpha)$ is decreasing in α. To make sure for any $s < \bar{s}$, $\alpha(\boldsymbol{k})$ satisfies

$\alpha(\boldsymbol{k}) > 1$, we require $\bar{\alpha} > 1$. Therefore,

$$\rho = \sum_{i=1}^{N} p_i \left(1 - \frac{1}{\sum_{j=0}^{1+\lfloor d_i \rfloor} \bar{\alpha}^j}\right) > \rho = \sum_{i=1}^{N} p_i \left(1 - \frac{1}{2 + \lfloor d_i \rfloor}\right).$$

This condition implies λ should be high enough in part (v).

Step 4: Finally, we show $\bar{s}$ is increasing in d_N for integer d_N. From the definition of $\bar{s}$ in (B.17), it suffices to show that $\eta(\alpha(d_N + 1)) > \eta(\alpha(d_N))$, where $\alpha(x)$ solves $\rho = \sum_{i=1}^{N-1} p_i \left(1 - \frac{1}{\sum_{j=0}^{1+\lfloor d_i \rfloor} \alpha(x)^j}\right) + p_N \left(1 - \frac{1}{\sum_{j=0}^{1+x} \alpha(x)^j}\right)$. From the definition of $\alpha(x)$, we have $\alpha(d_N + 1) < \alpha(d_N)$. Since $\eta(\alpha)$ is decreasing in α for fixed $d_1, \ldots, d_N$, it suffices to show

$$\frac{\sum_{j=0}^{d_N+1} [\alpha(d_N + 1)]^j \, (d_N + 1 - j)}{\sum_{l=0}^{2+d_N} [\alpha(d_N + 1)]^l} - \frac{\sum_{j=0}^{d} [\alpha(d_N)]^j \, (d_N - j)}{\sum_{l=0}^{1+d_N} [\alpha(d_N)]^l} > 0.$$

For any positive integer d,

$$\begin{aligned} &\frac{\sum_{j=0}^{d+1} [\alpha(d + 1)]^j \, (d + 1 - j)}{\sum_{l=0}^{2+d} [\alpha(d + 1)]^l} - \frac{\sum_{j=0}^{d} [\alpha(d)]^j \, (d - j)}{\sum_{l=0}^{1+d} [\alpha(d)]^l} \\ &> \frac{\sum_{j=0}^{d+1} [\alpha(d)]^j \, (d + 1 - j)}{\sum_{l=0}^{2+d} [\alpha(d)]^l} - \frac{\sum_{j=0}^{d} [\alpha(d)]^j \, (d - j)}{\sum_{l=0}^{1+d} [\alpha(d)]^l}. \end{aligned}$$

Therefore, it suffices to show $\frac{\sum_{j=0}^{d+1} \alpha^j (d+1-j)}{\sum_{l=0}^{2+d} \alpha^l} - \frac{\sum_{j=0}^{d} \alpha^j (d-j)}{\sum_{l=0}^{1+d} \alpha^l} > 0$ for any $\alpha > 0$. This follows

from the following derivation:

$$\begin{aligned}
&\sum_{l=0}^{1+d}\alpha^l\sum_{j=0}^{d+1}\alpha^j\left(d+1-j\right)-\sum_{l=0}^{2+d}\alpha^l\sum_{j=0}^{d}\alpha^j\left(d-j\right)\\
=&\sum_{l=0}^{1+d}\alpha^l\left[\sum_{j=0}^{d}\alpha^j\left(d-j\right)+\sum_{j=0}^{d}\alpha^j\right]-\left(\sum_{l=0}^{1+d}\alpha^l+\alpha^{2+d}\right)\sum_{j=0}^{d}\alpha^j\left(d-j\right)\\
=&\sum_{l=0}^{1+d}\alpha^l\sum_{j=0}^{d}\alpha^j-\alpha^{2+d}\sum_{j=0}^{d}\alpha^j\left(d-j\right)\\
=&\frac{1-\alpha^{1+d}-\left(1+d\right)\alpha^{2+d}\left(1-\alpha\right)}{\left(1-\alpha\right)^2}\\
=&\frac{\left(1-\alpha\right)\left[\sum_{j=0}^{d}\left(\alpha^j-\alpha^{2+d}\right)\right]}{\left(1-\alpha\right)^2}>0.
\end{aligned}$$

This completes the proof that $\bar{s}$ is increasing in d_N for integer d_N. □

Proof of Proposition 2.3. The proof is carried out in three steps. Parts (i) and (ii) are proved in Step 1; Step 2 and 3 prove part (iii). In the proof, we suppress the superscript $*$ to denote equilibrium for notational convenience.

Step 1: $\bar{U}(\boldsymbol{k})\geq\bar{U}(\boldsymbol{k'})$.

Refer to $F(u)$ defined in Equation (B.1). Since $\boldsymbol{k}$ is the Pareto-dominant equilibrium under s, this implies for any $u>\bar{U}(\boldsymbol{k})$, $F(u)<u$. By the definition of $F(u)$, since $s<s'$, it is immediate that $F(u;s)>F(u;s'),\forall u$, where $F(u;s)$ is the notation for $F(u)$ under model primitive s. Hence, $F(u;s')<u$ for any $u>\bar{U}(\boldsymbol{k})$, which implies that any $u>\bar{U}(\boldsymbol{k})$ does not constitute an equilibrium under s'. Thus, the Pareto dominant equilibrium $\boldsymbol{k'}$ under s' must satisfy $\bar{U}(\boldsymbol{k'})\leq\bar{U}(\boldsymbol{k})$. Note that this prove is general and does not rely on $\boldsymbol{k}$ being a pure strategy.

If both $\boldsymbol{k'}$ and $\boldsymbol{k}$ are pure strategy equilibria, then $k_i=(\lfloor(V_i-U(\boldsymbol{k}))\frac{\mu}{c}\rfloor)^+,\ \forall i$, this implies $\boldsymbol{k}\leq\boldsymbol{k'}$ component-wise in part (i). By Equation (2.2), $\boldsymbol{k}\leq\boldsymbol{k'}$ component-wise implies $\alpha(\boldsymbol{k})\geq\alpha(\boldsymbol{k'})$. Since $ES(\boldsymbol{k})=\alpha(\boldsymbol{k})/\rho$, it follows that $ES(\boldsymbol{k})\geq ES(\boldsymbol{k'})$ in part (ii).

Now we show part (iii). Let $\Theta(\boldsymbol{k}, k_h, k_l) \triangleq \frac{\sum_{j=0}^{k_h} \alpha(\boldsymbol{k})^j}{\sum_{j=0}^{k_l} \alpha(\boldsymbol{k})^j}$, where k_h and k_l are the h-th and l-th component of $\boldsymbol{k}$, respectively, i.e., k_h (k_l) is the search threshold for quality level h (l) in the search threshold vector $\boldsymbol{k}$.

Step 2: $\Theta(\boldsymbol{k}, k_h, k_l) \geq \Theta(\boldsymbol{k'}, k'_h, k'_l)$ for any given $1 \leq l < h \leq N$, where $k'_h - k'_l = k_h - k_l \equiv d$.

We first show $\frac{\sum_{j=0}^{k_h} \alpha(\boldsymbol{k})^j}{\sum_{j=0}^{k_l} \alpha(\boldsymbol{k})^j} \geq \frac{\sum_{j=0}^{k'_h} \alpha(\boldsymbol{k})^j}{\sum_{j=0}^{k'_l} \alpha(\boldsymbol{k})^j}$. It is equivalent to showing

$$\sum_{j=0}^{k_h} \alpha(\boldsymbol{k})^j \sum_{j=0}^{k'_l} \alpha(\boldsymbol{k})^j - \sum_{j=0}^{k'_h} \alpha(\boldsymbol{k})^j \sum_{j=0}^{k_l} \alpha(\boldsymbol{k})^j \geq 0. \tag{B.18}$$

- If $\alpha(\boldsymbol{k}) = 1$, LHS of (B.18) simplifies to

$$LHS = (1 + k_h)(1 + k'_l) - (1 + k'_h)(1 + k_l) = k_h k'_l + k_h + k'_l - k'_h k_l - k'_h - k_l.$$

Since $k_h + k'_l = k'_h + k_l$, $LHS = k_h k'_l - k'_h k_l = k_h(k'_h - d) - k'_h(k_h - d) = d(k'_h - k_h) \geq 0$.

- If $\alpha(\boldsymbol{k}) \neq 1$, (B.18) simplifies to

$$\left[1 - \alpha(\boldsymbol{k})^{k_h+1}\right] \left[1 - \alpha(\boldsymbol{k})^{k'_l+1}\right] - \left[1 - \alpha(\boldsymbol{k})^{k'_h+1}\right] \left[1 - \alpha(\boldsymbol{k})^{k_l+1}\right] \geq 0.$$

This is equivalent to showing

$$\begin{aligned} &\alpha(\boldsymbol{k})^{k'_h+1} + \alpha(\boldsymbol{k})^{k_l+1} - \alpha(\boldsymbol{k})^{k_h+1} - \alpha(\boldsymbol{k})^{k'_l+1} \geq 0, \\ &\Leftrightarrow [\alpha(\boldsymbol{k})^{k_l+1} - \alpha(\boldsymbol{k})^{k_h+1}][1 - \alpha(\boldsymbol{k})^{k'_l-k_l}] \geq 0. \end{aligned}$$

The last step uses the condition $k'_l - k_l = k'_h - k_h$. This last inequality holds since $k_h \geq k_l$ and $k'_l \geq k_l$. Hence, $\frac{\sum_{j=0}^{k_h} \alpha(\boldsymbol{k})^j}{\sum_{j=0}^{k_l} \alpha(\boldsymbol{k})^j} \geq \frac{\sum_{j=0}^{k'_h} \alpha(\boldsymbol{k})^j}{\sum_{j=0}^{k'_l} \alpha(\boldsymbol{k})^j}$.

Since $\alpha(\boldsymbol{k}) \geq \alpha(\boldsymbol{k'})$, and $\frac{\sum_{j=0}^{k'_h} \alpha(\boldsymbol{k})^j}{\sum_{j=0}^{k'_l} \alpha(\boldsymbol{k})^j}$ is increasing in $\alpha(k)$, therefore, $\frac{\sum_{j=0}^{k'_h} \alpha(\boldsymbol{k})^j}{\sum_{j=0}^{k'_l} \alpha(\boldsymbol{k})^j} \geq \frac{\sum_{j=0}^{k'_h} \alpha(\boldsymbol{k'})^j}{\sum_{j=0}^{k'_l} \alpha(\boldsymbol{k'})^j}$. Consequently,

$$\frac{\sum_{j=0}^{k_h} \alpha(\boldsymbol{k})^j}{\sum_{j=0}^{k_l} \alpha(\boldsymbol{k})^j} \geq \frac{\sum_{j=0}^{k'_h} \alpha(\boldsymbol{k'})^j}{\sum_{j=0}^{k'_l} \alpha(\boldsymbol{k'})^j}, \text{ or } \Theta(\boldsymbol{k}, k_h, k_l) \geq \Theta(\boldsymbol{k'}, k'_h, k'_l).$$

Let random variable $V(\boldsymbol{k})$ denote the quality a customer obtains in equilibrium $\boldsymbol{k}$ with $\mathbb{P}(V = V_i) = \phi_i(\boldsymbol{k}),\ i = 1, \ldots, N$.

Step 3: $V(\boldsymbol{k'}) \leq_{st} V(\boldsymbol{k})$.

From Step 2, given $i \in \{1, \ldots, N\}$, for any $l \leq i < h$, we have

$$\frac{\frac{p_l}{\sum_{j=0}^{k_l} \alpha(\boldsymbol{k})^j}}{\frac{p_h}{\sum_{j=0}^{k_h} \alpha(\boldsymbol{k})^j}} \geq \frac{\frac{p_l}{\sum_{j=0}^{k'_l} \alpha(\boldsymbol{k'})^j}}{\frac{p_h}{\sum_{j=0}^{k'_h} \alpha(\boldsymbol{k'})^j}}.$$

Thus, for any given $h > i$, summing i inequalities over $l = 0, \ldots, i$ gives

$$\frac{\sum_{l=0}^{i} \frac{p_l}{\sum_{j=0}^{k_l} \alpha(\boldsymbol{k})^j}}{\frac{p_h}{\sum_{j=0}^{k_h} \alpha(\boldsymbol{k})^j}} \geq \frac{\sum_{l=0}^{i} \frac{p_l}{\sum_{j=0}^{k'_l} \alpha(\boldsymbol{k'})^j}}{\frac{p_h}{\sum_{j=0}^{k'_h} \alpha(\boldsymbol{k'})^j}}.$$

Likewise, summing $N - i$ inequalities over $h = i + 1, \ldots, N$ gives

$$\frac{\sum_{l=0}^{i} \frac{p_l}{\sum_{j=0}^{k_l} \alpha(\boldsymbol{k})^j}}{\sum_{h=i+1}^{N} \frac{p_h}{\sum_{j=0}^{k_h} \alpha(\boldsymbol{k})^j}} \geq \frac{\sum_{l=0}^{i} \frac{p_l}{\sum_{j=0}^{k'_l} \alpha(\boldsymbol{k'})^j}}{\sum_{h=i+1}^{N} \frac{p_h}{\sum_{j=0}^{k'_h} \alpha(\boldsymbol{k'})^j}}.$$

From Equation (2.2), the sum of numerator and denominator is equal to $1 - \rho$ on both sides

of the inequality. Therefore,

$$\sum_{l=0}^{i} \frac{p_l}{\sum_{j=0}^{k_l} \alpha(\boldsymbol{k})^j} \geq \sum_{l=0}^{i} \frac{p_l}{\sum_{j=0}^{k'_l} \alpha(\boldsymbol{k'})^j}, \quad i \in \{1, \ldots, N\}. \tag{B.19}$$

Since $\sum_{l=0}^{i} \phi_l(\boldsymbol{k}) = \frac{1}{\rho} \sum_{l=0}^{i} p_l \left(1 - \frac{1}{\sum_{j=0}^{k_l} \alpha(\boldsymbol{k})^j}\right)$, thus from inequality (B.19), $\sum_{l=0}^{i} \phi_l(\boldsymbol{k}) \leq \sum_{l=0}^{i} \phi_l(\boldsymbol{k'}), \forall i$. Hence, $V(\boldsymbol{k'}) \leq_{st} V(\boldsymbol{k})$. It follows that $EV(\boldsymbol{k}) \geq EV(\boldsymbol{k'})$ in (iii).

Note that in the proof we implicitly assume $k_i > 0, \forall i$ for the sake of argument. It can be easily generalized to the case where search thresholds for some quality levels are zero. □

Proof of Proposition 2.4. Part (i) follows from Step 1 in the proof of Proposition 2.3. We shall prove Part (ii) by contradiction. Suppose that for any $\tilde{s} \in (s, s')$, the underly Pareto-dominant equilibrium is a pure strategy one. Thus, for any two search costs that result in the same pure strategy equilibrium, the average search rewards are equal. Let $\hat{s}$ be any jump point where the pure strategy equilibrium changes. By Proposition 2.3, $ES^*(\hat{s}-) \geq ES^*(\hat{s}+)$, where $ES^*(\hat{s}-)$ and $ES^*(\hat{s}+)$ are the average sample size at the left limit and right limit of $\hat{s}$. Also by Proposition 2.3, $\bar{U}^*(\hat{s}-) \geq \bar{U}^*(\hat{s}+)$. Hence, $\bar{U}^*(\hat{s-}) + sES^*(\hat{s}-) \geq \bar{U}^*(\hat{s}+) + sES^*(\hat{s}+)$, or $ER^*(\hat{s-}) \geq ER^*(\hat{s}+)$. This implies the average search reward is non-increasing in s. Hence, for $s < s'$, $ER^*(s) \geq ER^*(s')$. This leads to a contradiction, and therefore, there must exist $\tilde{s} \in (s, s')$ such that the underly Pareto-dominant equilibrium is a mixed strategy one. □

Proof of Proposition 2.5. The proof is analogous to that for Proposition 2.2. For com-

pleteness, we rewrite conditions (B.14) and (B.15) here

$$\rho = p_1 \frac{\kappa\alpha(\boldsymbol{k})}{1+\kappa\alpha(\boldsymbol{k})} + \sum_{i=2}^{N} p_i \left(1 - \frac{1}{\sum_{j=0}^{1+\lfloor d_i \rfloor} \alpha(\boldsymbol{k})^j}\right), \tag{B.20}$$

$$s = \frac{c}{\mu} \sum_{i=2}^{N} p_i \frac{\sum_{j=0}^{\lfloor d_i \rfloor} \alpha(\boldsymbol{k})^j \, (d_i - j)}{\sum_{l=0}^{1+\lfloor d_i \rfloor} \alpha(\boldsymbol{k})^l} \triangleq \frac{c}{\mu} \eta(\alpha(\boldsymbol{k})). \tag{B.21}$$

We impose conditions to ensure that the search thresholds $(\kappa, 1+\lfloor d_2 \rfloor, \ldots, 1+\lfloor d_N \rfloor)$ constitutes a Pareto dominant equilibrium. Note from (B.21) that since s is now fixed, $\alpha(\boldsymbol{k})$ does not change with ρ. Hence, $\bar{\lambda}$ is determined by setting $\kappa = 1$.

$$\bar{\lambda} = \mu \sum_{i=1}^{N} p_i \left(1 - \frac{1}{\sum_{j=0}^{1+\lfloor d_i \rfloor} \tilde{\alpha} j}\right) \quad \text{where } \tilde{\alpha} \text{ solves } s = \frac{c}{\mu}\eta(\tilde{\alpha}). \tag{B.22}$$

Since $\eta(\alpha)$ is decreasing in α, this equation admits a positive root $\tilde{\alpha}$ if and only if $s < \frac{c}{\mu}\eta(0) = \frac{c}{\mu}\sum_{i=2}^{N} p_i d_i$.

Let $\underline{\lambda} = \mu(1-p_1)$. Thus, if $\bar{\lambda} > \underline{\lambda}$, then search thresholds $(\kappa, 1+\lfloor d_2 \rfloor, \ldots, 1+\lfloor d_N \rfloor)$ will be Pareto dominant for any $\lambda \in (\underline{\lambda}, \bar{\lambda})$. This is because we have $\lambda > \mu(1-p_1)$, which ensures any equilibrium with a higher $\bar{U}$ would be instable (This is a similar argument as in Proposition 2.2). From (B.22), $\bar{\lambda}$ is increasing in $\tilde{\alpha}$ and $\tilde{\alpha}$ is decreasing in s. Hence, a small enough s would guarantee $\bar{\lambda} > \underline{\lambda}$. This also implies that $\bar{\lambda}$ increases as s decreases. This proves the conditions imposed for part (ii) to hold.

Moreover, we can also show that $\bar{\lambda}$ is increasing in integer d_N. This is shown by showing $\tilde{\alpha}(\boldsymbol{k})$ increases in integer d_N (see (B.22)). Recall from the proof of Proposition 2.2), we have shown that $\eta(\alpha; d_{N+1}) > \eta(\alpha; d_N)$ for any α. Since $\eta(\alpha(d_{N+1}); d_N + 1) = \eta(\alpha(d_N); d_N) = s\mu/c$ and $\eta(\alpha)$ is decreasing in α, it must be that $\alpha(d_{N+1}) > \alpha(d_N)$. Therefore, $\bar{\lambda}$ is increasing in integer d_N.

As arrival rate λ decreases, $ES(\boldsymbol{k}) = \alpha(\boldsymbol{k})/\rho$ increases, which proves part (i). From (B.20), since $\alpha(\boldsymbol{k})$ is unchanged, decreasing ρ on the LHS decreases κ. Therefore, κ is an

increasing function of ρ. Furthermore, for any $i \geq 2$, $\phi_i(\boldsymbol{k}) = p_i \left(1 - \frac{1}{\sum_{j=0}^{1+\lfloor d_i \rfloor} \alpha(\boldsymbol{k})^j}\right)/\rho$ increases as ρ decreases. Hence part (iii) follows that $EV(\boldsymbol{k}) = \sum_{i=1}^{N} \phi_i(\boldsymbol{k}) V_i$ increases. Since for $\lfloor d_N \rfloor > 1$, $EW_i(\boldsymbol{k}) \geq EW_1(\boldsymbol{k}) = 1/\mu$ with $EW_N(\boldsymbol{k}) > EW_1(\boldsymbol{k})$ and $EW_i(\boldsymbol{k})$ for all i is unchanged as $\alpha(\boldsymbol{k})$ is unchanged, $EW(\boldsymbol{k}) = \sum_{i=1}^{N} \phi_i(\boldsymbol{k}) EW_i(\boldsymbol{k})$ increases as λ decreases (part (iv)). Since $ER = \bar{U} + sES$, $\bar{U}$ is constant and ES increases, part (v) follows. □

Proof of Proposition 2.6. The proof is carried out in three steps. Parts (i) is proved in Step 1; Parts (ii-a) and (ii-b) are proved in Step 2 and 3. In the proof, we suppress the superscript $*$ to denote equilibrium for notational convenience. Since $\boldsymbol{k}^*$ is unchanged, we also drop the dependence of system metrics on $\boldsymbol{k}^*$.

Step 1: ES decreases. By the implicit function theorem, from Equation (2.2), decreasing λ decreases α for fixed $\boldsymbol{k}$. Since $ES = \left(1 - \sum_{i=1}^{N} \frac{p_i \alpha^{k_i}}{\sum_{j=0}^{k_i} \alpha^j}\right)^{-1}$ is increasing in α, part (i) follows. Since EW_i is increasing in α, $\forall i$, it follows that EW_i decreases as λ decreases.

Step 2: $\varphi(\alpha)$ is increasing in α for $\alpha \in (0,1)$ and decreasing in α for $\alpha > 1$, where, for given integers $k_h > k_l$, $\varphi(\alpha)$ is defined as

$$\varphi(\alpha) \triangleq \frac{1 - \frac{\alpha^{k_h}}{\sum_{j=0}^{k_h} \alpha^j}}{1 - \frac{\alpha^{k_l}}{\sum_{j=0}^{k_l} \alpha^j}}.$$

$$\begin{aligned}
\varphi(\alpha) &= \frac{1 - \frac{\alpha^{k_h}}{\sum_{j=0}^{k_h} \alpha^j}}{1 - \frac{\alpha^{k_l}}{\sum_{j=0}^{k_l} \alpha^j}} = \frac{\frac{\sum_{j=0}^{k_h-1} \alpha^j}{\sum_{j=0}^{k_h} \alpha^j}}{\frac{\sum_{j=0}^{k_l-1} \alpha^j}{\sum_{j=0}^{k_l} \alpha^j}} = \frac{\alpha \frac{\sum_{j=0}^{k_h-1} \alpha^j}{\sum_{j=0}^{k_h} \alpha^j}}{\alpha \frac{\sum_{j=0}^{k_l-1} \alpha^j}{\sum_{j=0}^{k_l} \alpha^j}} = \frac{\frac{\sum_{j=0}^{k_h-1} \alpha^{j+1}}{\sum_{j=0}^{k_h} \alpha^j}}{\frac{\sum_{j=0}^{k_l-1} \alpha^{j+1}}{\sum_{j=0}^{k_l} \alpha^j}} = \frac{\frac{\sum_{j=1}^{k_h} \alpha^j}{\sum_{j=0}^{k_h} \alpha^j}}{\frac{\sum_{j=1}^{k_l} \alpha^j}{\sum_{j=0}^{k_l} \alpha^j}} = \frac{1 - \frac{1}{\sum_{j=0}^{k_h} \alpha^j}}{1 - \frac{1}{\sum_{j=0}^{k_l} \alpha^j}} \\
&= \frac{1 - \frac{1-\alpha}{1-\alpha^{k_h+1}}}{1 - \frac{1-\alpha}{1-\alpha^{k_l+1}}} = \frac{1-\alpha^{k_h}}{1-\alpha^{k_l}} \cdot \frac{1-\alpha^{k_l+1}}{1-\alpha^{k_h+1}}.
\end{aligned}$$

Observe that $\varphi(\alpha) = \varphi(\frac{1}{\alpha})$. Hence, it suffices to show that $\varphi(\alpha)$ is increasing in α for $\alpha \in (0,1)$. It further suffices to show $\ln(\varphi(\alpha))$ is increasing in α for $\alpha \in (0,1)$. Define $g(\alpha, k)$ such that

$$\ln(\varphi(\alpha)) = \int_{k_l}^{k_h} g(\alpha, k) dk.$$

Thus,

$$\begin{aligned} g(\alpha, k) &= \lim_{\epsilon \to 0} \frac{\partial}{\partial \epsilon} \left[\ln(1 - \alpha^{k+\epsilon}) + \ln(1 - \alpha^{k+1}) - \ln(1 - \alpha^k) - \ln(1 - \alpha^{k+1+\epsilon}) \right] \\ &= -\frac{\alpha^k \ln \alpha}{1 - \alpha^k} + \frac{\alpha^{k+1} \ln \alpha}{1 - \alpha^{k+1}} \\ &= \left(\frac{1}{1 - \alpha^{k+1}} - \frac{1}{1 - \alpha^k} \right) \ln \alpha \end{aligned}$$

It suffices to show $g(\alpha, k)$ is increasing in $\alpha \in (0,1)\ \forall\ k > 0$. Define $h(\alpha, k)$ such that

$$g(\alpha, k) = \int_{x=k}^{k+1} h(\alpha, x) dx.$$

Thus,

$$h(\alpha, x) \triangleq \lim_{\epsilon \to 0} \frac{\partial}{\partial \epsilon} \left[\left(\frac{1}{1 - \alpha^{x+\epsilon}} - \frac{1}{1 - \alpha^x} \right) \ln \alpha \right] = (\ln \alpha)^2 \frac{\alpha^x}{(1 - \alpha^x)^2}.$$

It suffices to show $h(\alpha, x)$ is increasing in $\alpha \in (0,1)\ \forall\ x > 0$. Let $b = \alpha^x$. Then given $x > 0$, b is increasing in α. Thus,

$$h(\alpha, x) = \frac{1}{x^2} \cdot \frac{b(\ln b)^2}{(1 - b)^2}$$

It suffices to $H(b) \triangleq \frac{b(\ln b)^2}{(1-b)^2}$ is increasing in $b \in (0,1)$. Taking the derivative yields

$$H'(b) = \frac{[2 - 2b + (1 + b) \ln b] \ln b}{(1 - b)^3}.$$

Since $\ln b < 0$ and $(1-b)^3 > 0$ for $b \in (0,1)$. It suffices to show $G(b) \triangleq 2 - 2b + (1+b) \ln b < 0$

for $b \in (0,1)$. Note that $G(1) = 0$. Hence it suffices to show $G'(b) > 0$. This is true since

$$G'(b) = \ln b - \left(1 - \frac{1}{b}\right) > 0.$$

Consequently, $\varphi(\alpha)$ is increasing in α for $\alpha \in (0,1)$ and decreasing in α for $\alpha > 1$.

Step 3: if $\lambda \in (0, \hat{\lambda})$, $V(\lambda) \leq_{lr} V(\lambda + \epsilon)$; if $\lambda \in (\hat{\lambda}, \mu)$, $V(\lambda + \epsilon) \leq_{lr} V(\lambda)$, where $\epsilon > 0$ is a small positive number and $V(\lambda)$ is the notation of the random variable indicating the quality customers obtain given λ.

Let k_h (k_l) be the search threshold for quality level h (l) in the search threshold vector $\boldsymbol{k}$. By definition, $\phi_h/\phi_l = \varphi(\alpha)\frac{p_h}{p_l}$, $h > l$. Note that $\alpha = 1$ for $\lambda = \hat{\lambda}$ and α is increasing in λ by Step 1. Thus, by Step 2, $\phi_h(\lambda)/\phi_l(\lambda)$ is weakly increasing (decreasing) in λ if $\lambda \leq \hat{\lambda}$ ($\lambda > \hat{\lambda}$). Thus, if $\lambda \leq \hat{\lambda}$, $\phi_h(\lambda)/\phi_l(\lambda) \leq \phi_h(\lambda+\epsilon)/\phi_l(\lambda+\epsilon)$. Shuffling terms yields $\phi_h(\lambda)/\phi_h(\lambda+\epsilon) \leq \phi_l(\lambda)/\phi_l(\lambda+\epsilon)$. This implies $\phi_i(\lambda+\epsilon)/\phi_i(\lambda)$ is weakly increasing in i. Therefore, $V(\lambda) \leq_{lr} V(\lambda+\epsilon)$. We can analogously prove $V(\lambda+\epsilon) \leq_{lr} V(\lambda)$ for $\lambda \in (\hat{\lambda}, \mu)$.

The signs of $\frac{\partial}{\partial\lambda}EV$ in parts (ii-a) and (ii-b) immediately follow.

When $\lambda \in (0, \hat{\lambda})$,

$$EW(\lambda+\epsilon) = \sum_{i=1}^{N} \phi_i(\lambda+\epsilon) EW_i(\lambda+\epsilon) \geq \sum_{i=1}^{N} \phi_i(\lambda) EW_i(\lambda+\epsilon) \geq \sum_{i=1}^{N} \phi_i(\lambda) EW_i(\lambda) = EW(\lambda).$$

The first inequality is by Step 3. The second equality is by the result that EW_i decreases as λ decreases in Step 1. This shows EW decreases as λ decreases for $\lambda \in (0, \hat{\lambda})$ in part (ii-a).

□

B.3 Five Principles Hold under Homogeneous Quality $N = 1$

- Principle 1: reducing search costs leads to sampling more servers before joining.
- Principle 2: reducing search costs strictly increases customer welfare.

- Principle 3: reducing search costs weakly increases the average search reward.
- Principle 4: reducing the arrival rate (a) reduces the average waiting time, and, hence, (b) strictly increases the average search reward and (c) strictly increases individual customer welfare.
- Principle 5: controlling for the arrival rate, sampling more servers before joining (a) reduces the average waiting time, and, hence, (b) strictly increases the average search reward and (c) strictly increases individual customer welfare.

Principle 1 follows from Proposition 2.3 and Theorem 2.2. Principle 2 follows from Proposition 2.4 and Theorem 2.2. Specifically, customer welfare cannot be flat in search costs because mixed strategy equilibria are not Pareto dominant. Principle 3 follows from Proposition 2.4 and Theorem 2.2 (see the discussion after Proposition 2.4). Principle 5 is implied by Principle 1, 2 and 3 since there is quality differentiation and the increase in average search reward implies the reduction in the average waiting time. Now, we prove Principle 4.

Proof of Principle 4. We first prove that part (a) EW decreases as λ decreases. This is divided into two cases. We first prove the case when the equilibrium search threshold k is not changed a later the case when it is changed. Consider the case in which k is not changed. Note that $EW = \frac{\sum_{j=0}^{k-1}(j+1)\alpha(k)^j}{\sum_{j=0}^{k-1}\alpha(k)^j}$ is increasing in $\alpha(k)$. Since $\alpha(k)$ decreases as ρ decreases (from the proof of Proposition 2.6), it follows that EW decreases.

Now we prove the case when the equilibrium search threshold is changed. Let $\lambda < \lambda'$. Since $EW(k;\lambda) < EW(k;\lambda')$ for any k, $\lambda EW(k;\lambda) < \lambda EW(k;\lambda') < \lambda' EW(k;\lambda')$. By Little's law, $EQ(k;\lambda) < EQ(k;\lambda')$. Let k^* be the Pareto-dominant equilibrium under λ and $k^{*\prime}$ be any pure strategy equilibrium under λ'. We next show $k^* \leq k^{*\prime}$. From Theorem 2.2, $k^* \leq 1 + \frac{s\mu}{c} + EQ(k^*;\lambda) \leq k^* + 1$, and since k^* is the Pareto-dominant equilibrium, $1+\frac{s\mu}{c}+EQ(k;\lambda) > k+1$ for $k < k^*$. Since $EQ(k;\lambda) < EQ(k;\lambda')$, $1+\frac{s\mu}{c}+EQ(k;\lambda') > k+1$ for $k < k^*$. Thus, any $k < k^*$ cannot be an equilibrium under λ'. Therefore, $k^* \leq k^{*\prime}$. For

any $k \leq k'$, $EW(k;\lambda) < EW(k;\lambda') \leq EW(k';\lambda')$. Therefore, $EW(k^*;\lambda) < EW(k^{*\prime};\lambda')$. This proves part (a).

Part (b) is equivalent with part (a) since $N = 1$. For part (c), if k is not changed, EW decreases and ES decreases (see Proposition 2.6), hence individual customer welfare improves. If reducing λ leads to a jump in k, k always gets smaller, which implies at those jump points, individual customer welfare also improves. □

B.4 Numerical Studies

To gain a broader perspective about when the average waiting time rises as either the search cost or arrival rate fall, we run more numerical trials under various parameter settings. For all the results reported, we consider two quality levels, $N = 2$, $\mu = 1$, and $c = 1$.

B.4.1 Impact of Search Cost Reduction

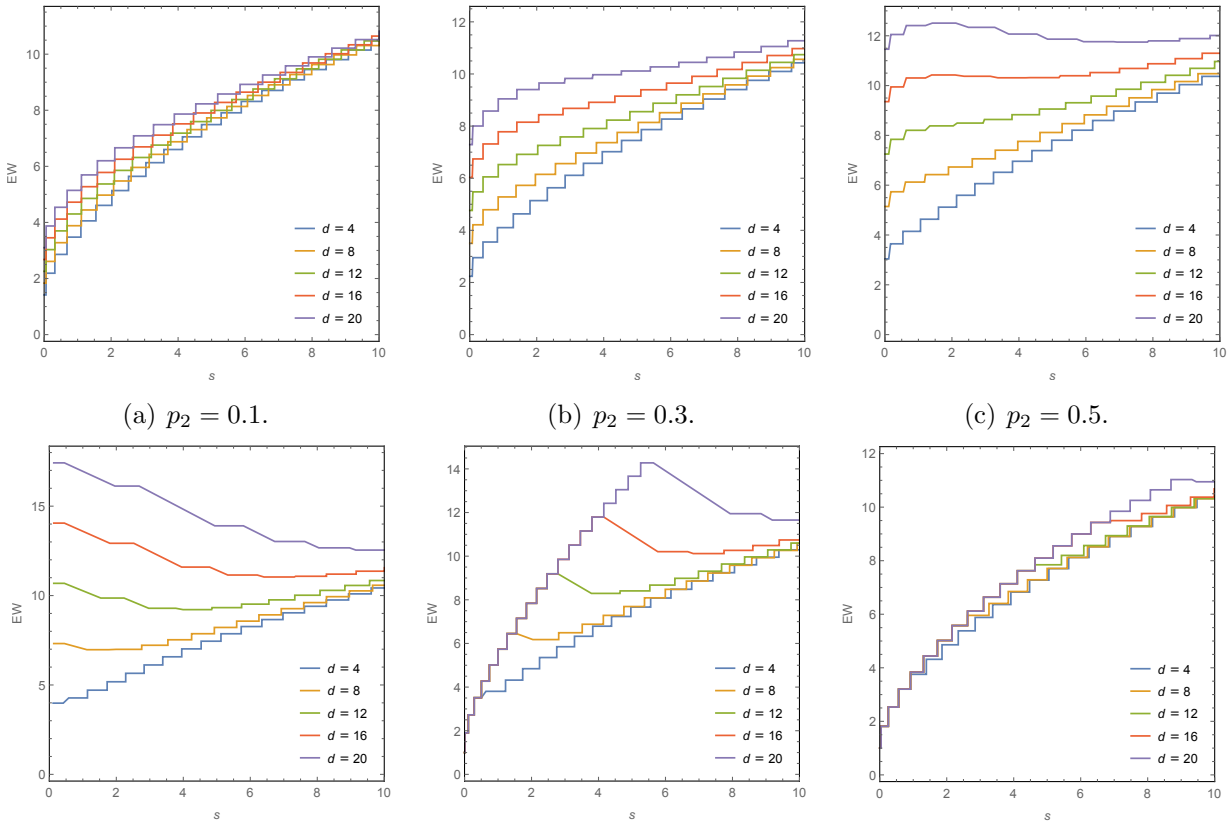
Figure B.2 shows the impact of search cost on the average waiting time. We have the following observations:

- *Observation 1*: Under mild quality differences, e.g., $d = 4$, the average waiting time is almost always monotone in search cost regardless of p_2, suggesting that sufficient quality differences across servers are pivotal to the emergence of the non-monotone comparative statics.[1]

- *Observation 2*: For fixed p_2, as the quality difference d grows, the negative relationship between search costs and the average waiting time becomes more pronounced. For example, when $p_2 = 0.8$, $d = 20$, the average waiting time is downward sloping in search cost s for all s considered in the experiment (up to 10).

1. When d is small, the interval $(0, \bar{s})$ defined in Proposition 2.2 within which the average waiting time is decreasing in search cost is too small to be adequately shown in the figures.

Figure B.2: The average waiting time versus search costs under different p_2.



(a) $p_2 = 0.1$. (b) $p_2 = 0.3$. (c) $p_2 = 0.5$.

(d) $p_2 = 0.8$. (e) $p_2 = 0.96$. (f) $p_2 = 0.99$.

Note. $N = 2$, $\mu = 1$, $\lambda = 0.95$, $c = 1$. $d = (V_2 - V_1)\mu/c$.

- *Observation 3*: The average waiting time rises as search cost falls mostly at small search cost levels. Interestingly, the average waiting time is not always unimodal in search costs in general. For instance, when $p_2 = 0.5$, for either $d = 16$ or $d = 20$, the average waiting time initially increases, then decreases, and eventually increases with search costs. An exception that defies this "small threshold property" is the case $p_2 = 0.96$, where for large d, the phenomenon occurs when search costs are relatively large. This is because in this case, at small search cost levels, all customers join high quality servers, which essentially corresponds to the single quality level benchmark. Only at relatively large search cost levels (around 10) do low quality servers enjoy positive throughput. At even larger search cost levels (not shown on the figure), the

curve will, once again, be upward sloping.

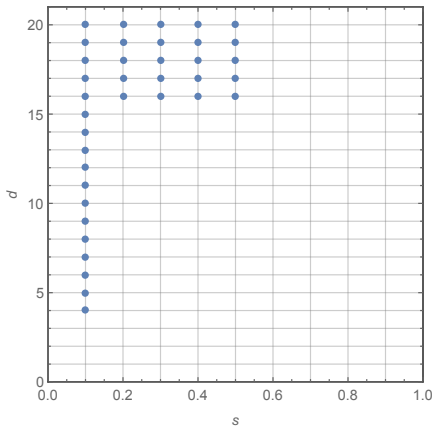
- *Observation 4*: Quality heterogeneity, proxied by p_2, is critical to the direction of the comparative statics. If p_2 is small, e.g., $p_2 = 0.1$ and $p_2 = 0.3$, we do not observe non-monotonicity even for large d under the parameters considered. When p_2 is large, e.g., $p_2 = 0.99$, there is not much non-monotone behavior either (with the only exceptional case where $d = 20$ and s is close to 10). This is not surprising since the quality substitution effect will be negligible when p_2 is close to either end of the $[0, 1]$ interval. An interesting observation is the asymmetry effect of p_2. The negative relationship between search costs and the average waiting time is most pervasive (among the four sub-figures) when $p_2 = 0.8$, not when the quality distribution is most heterogeneous (with the highest variance), i.e., when $p_2 = 0.5$. A skewed distribution toward high quality servers may be more conducive to the non-monotone behavior.

- *Observation 5*: For fixed p_2, the average waiting time tends to increase with d (with some notable exceptions when $p_2 = 0.1$). The equilibrium structure requires the threshold difference at different quality levels to be proportional to the quality difference itself; therefore, a larger d implies longer queues at high quality servers, and hence a longer wait. This manifests another detrimental aspect of quality differentiation.

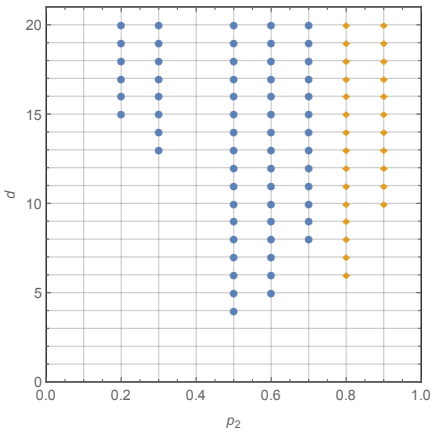
B.4.2 Impact of Arrival Rate Reduction

Figure B.3 shows the impact of arrival rate on the average waiting time. For the parameters at each grid point in the Figure, we first solve for the equilibrium search thresholds, then compute the derivative $\frac{\partial EW(\boldsymbol{k})}{\partial \lambda}$ given the search thresholds (note that the equilibrium search thresholds are unaffected if λ is changed by a small ϵ almost everywhere). We label each point with a filled circle (for a pure strategy equilibrium) or a diamond (for a mixed strategy equilibrium) if the derivate $\frac{\partial EW(\boldsymbol{k})}{\partial \lambda}$ under the given parameters is negative. This allows us to numerically investigate when a lower arrival rate would lead to a higher average waiting

Figure B.3: The average waiting time rises as the arrival rate falls when equilibrium search thresholds do not change.



(a) Search cost.



(b) The fraction of high quality servers.

Note. The filled circles (pure strategy equilibria) and diamonds (mixed strategy equilibria) indicate points where $\frac{\partial EW(\boldsymbol{k})}{\partial \lambda} < 0$. $p_2 = 0.5$, $\lambda = 0.95$, $s = 0.1$ unless otherwise specified.

time (without threshold changes).

Figure B.3-(a) shows that a higher search cost requires a higher quality difference in order for the relationship between arrival rate and average waiting time to be negative. For example, when the search cost is increased from 0.1 to 0.2, the lowest quality difference needed jumps from 4 to 16. In fact, this search cost change leads to an equilibrium threshold change, which entails a larger quality difference to countervail larger thresholds. Figure B.3-(b) shows that heterogeneity in quality distribution, proxied by p_2, and the quality difference d are roughly substitutes: with greater heterogeneity (p_2 getting closer to 0.5), a lower quality difference tends to be warranted for the average waiting time to rise as the arrival rate falls. For instance, when $p_2 = 0.5$, only a minimum quality difference of 4 is required, the lowest among all the p_2 examined.

These numerical experiments verify our analytical insights for a broader set of parameters: as either the search cost or arrival rate falls, the average waiting time is most likely to rise when search costs are already low and quality differentiation is substantial, both of which stimulate customers' strategic search behavior.

Example B.2 and Table B.1 show an instance in which a decrease in the arrival rate

results in deterioration of the average search reward.

Example B.2. *Consider the following parameters:* $N = 2$, $\mu = 1$, $c = 1$, $s = 2.5$, $\boldsymbol{p} = (0.8, 0.2)$, $V_1 = 50$, $V_2 = 60$. *The system metrics under arrival rates 0.27 and 0.28 are shown in Table B.1. The equilibrium is* $\boldsymbol{k}^* = (1, 11)$ *under both arrival rates.*

Table B.1: System metrics versus arrival rate for Example B.2. As λ falls, ER^* deteriorates.

λ	ER^*	EV^*	EW^*
0.27	51.379	52.507	1.128
0.28	51.390	52.529	1.139

Albeit of theoretical interest, the pathological example shown in Table B.1 occurs when the system load ρ is small (see Proposition 2.6-(iii)), which may not be too relevant for practical purposes.

Example B.3 and Table B.2 show an instance in which reducing the arrival rate changes the (pure strategy) search thresholds and increases the average waiting time. It also illustrates that reducing both the arrival rate and search cost increases the average waiting time.

Example B.3. *Consider the following parameters:* $N = 2$, $\mu = 1$, $c = 1$, $\boldsymbol{p} = (0.6, 0.4)$, $d = 16$. *Table B.2 tabulates the average waiting time under the combination of two search costs, 3.7 and 5.6, and two arrival rates, 0.67 and 0.77. Fixing the arrival rate at either level, a fall in search cost from 5.6 to 3.7 leads to an rise in the average waiting time. Fixing the search cost at 5.6, a fall in the arrival rate from 0.77 to 0.67 leads to a fall in the average waiting time, which is intuitive. Fixing the search cost at 3.7, however, a fall in the arrival rate from 0.77 to 0.67 leads to rise in the average waiting time. When both the search cost and arrival rate fall, from* $(s, \lambda) = (5.6, 0.77)$ *to* $(3.7, 0.67)$, *the average waiting time rises.*

Table B.2: Impact of search cost and arrival rate on the average waiting time in the system.

		$\lambda = 0.67$	$\lambda = 0.77$
s =3.7	$EW(\boldsymbol{k})$	5.161	5.097
	$\boldsymbol{k}$	(1,17)	(2,18)
s =5.6	$EW(\boldsymbol{k})$	2.933	4.03
	$\boldsymbol{k}$	(2,18)	(3,19)

B.5 When Arrival Rate Reduction Changes Search Thresholds

We compare two pure strategy equilibrium under two different parameter settings. To gain analytical results, we look into a case in which $d = (V_2 - V_1)\mu/c \in \mathbb{N}_{>0}$ (which implies any pure strategy equilibrium with $k_1 \in \mathbb{N}_{>0}$ must be of the form $(k_1, k_1 + d)$). We also study the interplay of the search cost and arrival rate, and find conditions under which the average waiting time increases when both parameters fall.

Lemma B.2 lays the foundation for specifying these conditions. It compares the average waiting time in the system for two given adjacent thresholds that differ by 1 at both quality levels, each under a different arrival rate. The arrival rate λ^L associated with the lower thresholds $(k_1, k_1 + d)$ is smaller than the arrival rate λ^H associated with the higher thresholds.

Lemma B.2. *$N = 2$. Compare two given search thresholds $\boldsymbol{k}^L = (k_1, k_1 + d)$ and $\boldsymbol{k}^H = (k_1 + 1, k_1 + d + 1)$ under arrival rates λ^L and λ^H respectively, where $k_1 \in \mathbb{N}_{>0}$, $d \in \mathbb{N}_{>0}$, $\lambda^L < \lambda^H$ and $\lambda_L = \mu\left[1 - (\frac{p_1}{k_1+1} + \frac{p_2}{k_1+d+1})\right]$ and $\lambda_H = \mu\left[1 - (\frac{p_1}{k_1+2} + \frac{p_2}{k_1+d+2})\right]$. Thus, $EW(\boldsymbol{k}^L) > EW(\boldsymbol{k}^H)$ if both of the following conditions hold:*

(i) $d(d+1) > 2(1+k_1)^2$ and $\Delta(d, k_1) > 0$, where $\Delta(d, k_1)$ decreases in k_1 and increases in d for large enough d;

(ii) $p_2 \in (\underline{p}(d, k_1), \bar{p}(d, k_1))$, where the interval length, $\bar{p}(d, k_1) - \underline{p}(d, k_1)$, decreases in k and $\lim_{d\to\infty} \bar{p}(d, k_1) - \underline{p}(d, k_1) = 1$.

For tractability, we select λ^L and λ^H such that $\alpha(\boldsymbol{k}^L) = \alpha(\boldsymbol{k}^H) = 1$. This parameter choice (with $\lambda^L < \lambda^H$) significantly simplifies the expressions for the associated average

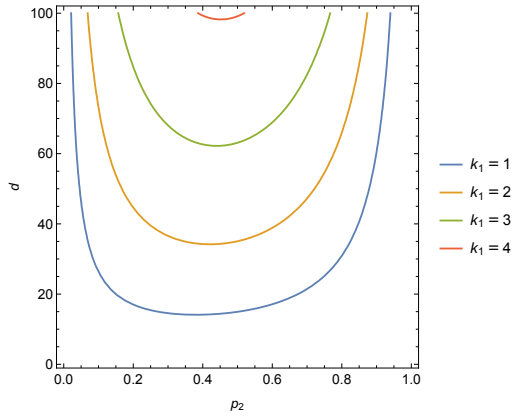
waiting times, which are, therefore, amenable to analytical comparison. According to Lemma B.2, two conditions must be satisfied in order for the average waiting time under lower search thresholds and a lower arrival rate to be greater than that under higher search thresholds and a higher arrival rate, a result that would arise if the upward pressure from quality substitution effect outstrips the downward pressure from the load balancing and system load effect.

The key insights from Condition (i) are that the difference between two quality levels, d, has to be large (an indicator of substantial quality differentiation), and that the two adjacent search thresholds under comparison have to be small (we give the formal expression of $\Delta(d, k_1)$ in Appendix B.2). The requirement of a small k_1 implies that the result $EW(\boldsymbol{k}^L) > EW(\boldsymbol{k}^H)$ is a "small threshold phenomenon." As we shall see in Proposition B.2, small search thresholds translate to small search cost levels in terms of model primitives.

Condition (ii) in Lemma B.2 says that the fraction of high quality servers, p_2, must lie within an intermediate range (the formal expressions of $(\underline{p}(d, k_1), \bar{p}(d, k_1))$ are given in Appendix B.2). Condition (ii) also suggests this requirement on p_2 is easier to satisfy when either k is small or d is sufficiently large, as in either case, the range for p_2 gets wider. Condition (ii) speaks to the importance of another aspect of quality differentiation: there must be a sufficient amount of quality heterogeneity. This requirement on heterogeneity will be less stringent if either search thresholds are small or the quality difference is large.

Figure B.4 illustrates Lemma B.2 by showing the interaction of d, k_1 and p_2 in driving the result $EW(\boldsymbol{k}^L) > EW(\boldsymbol{k}^H)$. For a given k_1, the curve partitions the parameter space (p_2, d) into two regions: $EW(\boldsymbol{k}^L) > EW(\boldsymbol{k}^H)$ if and only if the underlying (p_2, d) lies above the curve. For a given k_1, $EW(\boldsymbol{k}^L) > EW(\boldsymbol{k}^H)$ occurs only if d is greater than a certain value, and the p_2 lies within a certain interval that depends on d; as d gets larger, the interval also gets wider. If we increase the search threshold k_1, then d must be higher and p_2 must fall into a narrower interval. The $EW(\boldsymbol{k}^L) > EW(\boldsymbol{k}^H)$ region in the (p_2, d) space for lower k_1 envelops that for higher k_2, indicating more quality differentiation (both in terms of the

Figure B.4: The impact of d, k_1, p_2 on $EW(\boldsymbol{k}^L)$ and $EW(\boldsymbol{k}^H)$ under $\mu = 1$ and λ_L, λ_H given in Lemma B.2.



difference between quality levels and heterogeneity in distribution) is entailed for the same phenomenon to arise under higher search thresholds.

Note that Lemma B.2 compares the average waiting time under two *given* search thresholds. We do not specify under what conditions these search thresholds would emerge in equilibrium. Proposition B.2 provide these conditions.

Proposition B.2. *$N = 2$. If both conditions (i) and (ii) in Lemma B.2 hold,*

- *Case (i) if* $\frac{s\mu}{c} \in \left[\frac{k_1+dp_2}{2} - \frac{(d+2)k_1^2+2dp_2+k_1^2}{2(k_1+2)(d+k_1+2)}, \frac{k_1+dp_2}{2}\right]$, *then* $\boldsymbol{k}^L$ *and* $\boldsymbol{k}^H$ *are equilibria under* (s, λ_L) *and* (s, λ_H), *respectively;*
- *Case (ii) if* $\frac{s_L\mu}{c} \in \left[\frac{k_1+dp_2}{2} - \frac{k_1+dk_1+k_1^2+dp_2}{(1+k_1)(1+d+k_1)}, \frac{k_1+dp_2}{2} - \frac{(d+2)k_1^2+2dp_2+k_1^2}{2(k_1+2)(d+k_1+2)}\right]$ *and* $\frac{s_H\mu}{c} \in \left[\frac{k_1+dp_2}{2}, \frac{1+k_1+dp_2}{2}\right]$, *then* $s_L < s_H$; $\boldsymbol{k}^L$ *and* $\boldsymbol{k}^H$ *are equilibria under* (s_L, λ_L) *and* (s_H, λ_H), *respectively.*

$EW(\boldsymbol{k}^L) > EW(\boldsymbol{k}^H)$ in all Cases (i), (ii) and (iii).

Fixing the search cost, case (i) of Proposition B.2 shows that a lower arrival rate leads to a higher average waiting time in equilibrium under the specified conditions. Since $ES(\boldsymbol{k}) = \alpha(\boldsymbol{k})/\rho$ and $\alpha(\boldsymbol{k}) = 1$ under both λ_L and λ_H, it follows that customers search more here with a lower arrival rate, echoing the case in Proposition 2.5. Case (ii) of Proposition B.2

exemplifies a joint effect of search cost and arrival rate: even when both parameters fall, the average waiting time may still rise. This may be a case most relevant for practice, as different policy interventions are often implemented simultaneously. Note that since the conditions in Lemma B.2 hold for small k_1, the search cost intervals in Case (i) and (ii) imply that small search cost levels are required for the underlying results to apply.

B.5.1 Proof

Proof of Lemma B.2. In both equilibria, α=1. The expected waiting time is

$$EW^L = \frac{1}{\lambda_L}\left[\frac{(1-p_2)k_1}{2} + \frac{p_2(k_1+d)}{2}\right], \quad EW^H = \frac{1}{\lambda_H}\left[\frac{(1-p_2)(k_1+1)}{2} + \frac{p_2(k_1+d+1)}{2}\right].$$

$EW^L > EW^H$ if and only if

$$\frac{1}{\lambda_L}\left[\frac{(1-p_2)k_1}{2} + \frac{p_2(k_1+d)}{2}\right] - \frac{1}{\lambda_H}\left[\frac{(1-p_2)(k_1+1)}{2} + \frac{p_2(k_1+d+1)}{2}\right] > 0.$$

This is equivalent to

$$\lambda_H\left[(1-p_2)k_1 + p_2(k_1+d)\right] - \lambda_L\left[(1-p_2)(k_1+1) + p_2(k_1+d+1)\right] > 0.$$

Condition (i): substituting λ_H and λ_L gives

$$\begin{aligned}&\left[1 - \left(\frac{1-p_2}{k_1+2} + \frac{p_2}{k_1+d+2}\right)\right]\left[(1-p_2)k_1 + p_2(k_1+d)\right] \\ -&\left[1 - \left(\frac{1-p_2}{k_1+1} + \frac{p_2}{k_1+d+1}\right)\right]\left[(1-p_2)(k_1+1) + p_2(k_1+d+1)\right] > 0.\end{aligned}$$

Since $d, k_1 \geq 0$, this simplifies to $\tau(p_2, d, k_1) > 0$, where

$$\begin{aligned}\tau(p_2, d, k_1) \triangleq&(k+1)(d+k+1)(dp_2+k)\left(d(k+p_2+1) + k^2 + 3k + 2\right) \\ &- (k+2)(d+k+2)(dp_2+k+1)\left(dk + dp_2 + k^2 + k\right).\end{aligned}$$

By inspection, given d and k_1, $\tau(p_2, d, k_1)$ is a concave quadratic function of p_2. Write

$$\tau(p_2, d, k_1) = a(d, k_1)p_2^2 + b(d, k_1)p_2 + c(d, k_1),$$

where $a(d, k_1) = -d^2(3 + d + 2k_1)$, $b(d, k_1) = d(d + d^2 - 2(1 + k_1)^2)$ and $c(d, k_1) = -k_1(1 + k_1)[2 + d^2 + 3k_1 + k_1^2 + d(3 + 2k_1)]$. Moreover, $\tau(0, d, k_1) = c(d, k_1) < 0$ and $\tau(1, d, k_1) = -(2+k_1)(1+d+k_1)(2+d+k_1)(d+k_1+dk_1+k_1^2)+(1+k_1)(d+k_1)(1+d+k_1)(2+3k_1+k_1^2+d(2+k_1)) < 0$. Therefore, for $p_2 \in (0, 1)$, $\tau(p_2, d, k_1) > 0$ holds if and only if $b(d, k_1) > 0$ and $\Delta(d, k_1) = b(d, k_1)^2 - 4a(d, k_1)c(d, k_1) > 0$. $b(d, k_1) > 0$ gives $d(d + 1) > 2(1 + k_1)^2$ in (i).

$$\begin{aligned}\frac{\Delta(d, k_1)}{d^2} =& d^4 + d^3(2 - 4k_1 - 4k_1^2) - d^2(3 + 32k_1 + 44k_1^2 + 16k_1^3) - 4d(1 + 13k_1 + 27k_1^2 + 20k_1^3 + 5k_1^4) \\ & - 4(1 + k_1)^2(-1 + 4k_1 + 6k_1^2 + 2k_1^3) \\ =& 4 - 4d - 3d^2 + 2d^3 + d^4 + (-8 - 52d - 32d^2 - 4d^3)k_1 + (-52 - 108d - 44d^2 - 4d^3)k_1^2 \\ & + (-72 - 80d - 16d^2)k_1^3 + (-40 - 20d)k_1^4 - 8k_1^5.\end{aligned}$$

It is immediate that $\Delta(d, k_1)$ is decreasing in k_1. Fixing k_1, $\Delta(d, k_1)$ is increasing in d^6 if d gets large since d^6 will be the dominant term. This completes the proof for part (i).

Condition (ii): given d, k_1,
$p_2 \in (\underline{p}(d, k_1), \bar{p}(d, k_1)) \triangleq \left(\frac{-b(d,k_1)+\sqrt{\Delta(d,k_1)}}{2a(d,k_1)}, \frac{-b(d,k_1)-\sqrt{\Delta(d,k_1)}}{2a(d,k_1)} \right)$.

$$\bar{p}(d, k_1) - \underline{p}(d, k_1) = \frac{\sqrt{\Delta(d, k_1)}}{|a(d, k_1)|}.$$

The numerator is decreasing in k_1 and the denominator is increasing in k_1. Hence the interval length, $\bar{p}(d, k_1) - \underline{p}(d, k_1)$, is decreasing in k_1. As d tends to infinity, $\Delta(d, k_1)$ grows in the order of d^6, and therefore $\sqrt{\Delta(d, k_1)}$ grows in the order of d^3. $|a(d, k_1)|$ also grows in the order of d^3. Hence, as $d \to \infty$, $\bar{p}(d, k_1) - \underline{p}(d, k_1)$ approaches 1. □

Proof of Proposition B.2. In order for (k_1, k_1+d) to be an equilibrium, it must satisfy

$$k_1 + d \leq \gamma(k_1, k_1+d) \leq k_1 + d + 1.$$

Since $\alpha = 1$ and $\lambda_L/\mu = 1 - \frac{1-p_2}{k_1+1} - \frac{p_2}{1+k_1+d}$,

$$\gamma(k_1, k_1+d) = \frac{s(k_1)\mu}{c} \frac{1}{1 - \frac{1-p_2}{k_1+1} - \frac{p_2}{1+k_1+d}} + \frac{(1+d+k_1)(k_1+2dk_1+k_1^2+dp_2-dk_1p_2)}{2(k_1+dk_1+k_1^2+dp_2)}.$$

Collecting terms yields

$$\frac{s(k_1)\mu}{c} \in \left[\frac{k_1+dp_2}{2} - \frac{k_1+dk_1+k_1^2+dp_2}{(1+k_1)(1+d+k_1)}, \frac{k_1+dp_2}{2}\right] \triangleq [\delta_1, \delta_3].$$

Similarly, (k_1+1, k_1+d+1) to be an equilibrium, we must have

$$\frac{s(k_1+1)\mu}{c} \in \left[\frac{k_1+dp_2}{2} - \frac{(d+2)k_1^2+2dp_2+k_1^2}{2(k_1+2)(d+k_1+2)}, \frac{1+k_1+dp_2}{2}\right] \triangleq [\delta_2, \delta_4].$$

Note that $k_1 + dk_1 + k_1^2 + dp_2 - \frac{(d+2)k_1^2+2dp_2+k_1^2}{2} = \frac{k_1(k_1+d)}{2} > 0$. It is immediate that $\delta_1 < \delta_2 < \delta_3 < \delta_4$. Thus, Cases (i) and (ii) immediately follow. □

APPENDIX C

SUPPLEMENT TO CHAPTER 3

C.1 Technical Proofs

Proof of Lemma 3.1. 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 $\mathbb{P}[N = k]$.

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

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_2^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

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

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,

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

Proof of Corollary 3.1. This follows from the expressions derived in Lemma 3.1.

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

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 C.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 C.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 C.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.

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

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 3.1. 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.

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

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

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 (3.5a) and (3.5c) move in the same direction (either both increasing or both decreasing) as c_r changes. Subtracting (3.5c) from (3.5a) 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,

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

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

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

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

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

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

where

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

First, in (C.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 (C.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 (C.4) gives

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

Recall from (C.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 (3.5a) and (3.5c) 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 C.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

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

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

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

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 (3.5a) gives an equation solely in terms of β.

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

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 (C.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 (3.5a) and (3.5c) 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.

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

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

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

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 (C.1a) and (C.1b). This is equivalent to a system composed of (C.1a) and $h(\alpha^*,\beta^*)=0$. As in Step 2, plugging $\alpha^*=\alpha(\beta^*)$ from

$h(\alpha^*, \beta^*) = 0$ into (C.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 C.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:

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

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 3.1. 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 (C.7) in the proof of Proposition 3.1, 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 (3.4c), 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 3.1 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 3.1 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 (3.4a)-(3.4c) 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

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

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

Again, this system of equations always 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 any $c_r \in [\min\{c_r^l, c_r^m\}, \max\{c_r^l, c_r^m\}]$, there always exists a partial-referral equilibrium $(\kappa, 1)$. From equilibrium conditions (3.7a)-(3.7c), 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 3.2. This follows from the argument in Theorem 3.1. □

Proof of Corollary 3.3. 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 3.2. From Theorem 3.1, 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

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

where (α, β) solve (C.8) and (C.9). From (C.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 (C.10) can be rewritten as $c_r < m(\alpha)$. Let $\max_\alpha m(\alpha) = \tilde{c_r}$. Condition (C.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 (C.8), β is decreasing in Λ. This further implies $c/(\mu - \Lambda\beta)$ is increasing in Λ. Hence, from equation (C.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 (C.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 (C.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 3.3. 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 (3.7a)-(3.7b). Since base customers are indifferent to referrals, (3.7a) is equivalent to

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

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 C.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

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

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 3.1, 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 (3.5a),

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

From the proof of Proposition 3.1 (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 (C.12). The rest of the proof is similar to the partial-referral case. □

Next, we introduce Lemma C.2 that will be used in the proof of Theorem 3.2, Theorem 3.3, and Proposition 3.4.

Lemma C.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*

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

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

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

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

Proof of Lemma C.2. Part (i): Under FIFO, $\lambda^F = \Lambda\beta^F$, where β^F solves $\bar{V}\left(1-\beta^F\right) - c/(\mu - \Lambda\beta^F) = 0$. Under the partial-referral equilibrium, the throughput $\lambda^R = \Lambda\beta^R[1 + \kappa\alpha/(1-\alpha)]$, where (β, α^R) solves (C.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 $\bar{V}\left(1-\beta^R\right) - cW_2^B\left(\alpha, \beta^R, \kappa\right) = 0$ by (C.11), this implies

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

Plugging in $W_2^B\left(\alpha, \beta^R, \kappa\right) = \frac{\mu(1-\alpha)^2+\Lambda\beta^R\kappa\alpha}{[\mu(1-\alpha)-\Lambda\beta^R\kappa\alpha][\mu(1-\alpha)-\Lambda\beta^R(1+\kappa\alpha-\alpha)]}$ 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 $\bar{V}\left(1-\beta^R\right)-cW_2^B\left(\alpha,\beta^R,\kappa\right)=0$,

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

By substitution,

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

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 3.1. Also note that at c_r^l, $\bar{V}\left(1-\beta^R\right)-cW_2^B\left(\alpha,\beta^R\right)=0$. Therefore, the result follows from the same argument in part (i) by letting $\kappa=1$. □

Proof of Theorem 3.2. Combining Part (i) and (ii) in Lemma C.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 3.3, 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 $\rho \leq 1/2$, 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 $\rho \leq 1/2$, $\rho(-\alpha^3+4\alpha^2-5\alpha+2)-(1-\alpha)^2 < 0$, $\forall \alpha \in (0,1)$. 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 $\rho \leq 1/2$. □

Proof of Theorem 3.3. We prove this by finding a sufficient condition for condition (C.13) in Lemma C.2 when $\kappa = 0$. When $\kappa = 0$, condition (C.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 3.1)

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

which gives $\bar{V} > 5c/2\mu$. □

Proof of Proposition 3.4. When $c_r = 0$, by Corollary 3.2, 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 C.2, we would have a similar inequality as (C.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 C.2, this simplifies to

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

From Proposition 3.1, at $\bar{\Lambda}$, $\alpha = 0$. Also note from the proof of Lemma 3.1 (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 (C.16) gives $\bar{V} > 3c/\mu$. □

Proof of Proposition 3.5. 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 3.3, $\beta^R < \beta^F$, which implies $ICW^R < ICW^F$. This proves Part (i). Part (ii) immediately follows from Part (i). □

Proof of Proposition 3.6. We have shown in Proposition 3.2 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 3.2), decreasing the conversion rate and the need for priority. Therefore, customers would also not refer. □

Proof of Lemma 3.2. The proof is similar to that of Lemma 3.1. We prove a more general result with $\lambda = \Lambda\beta$, $q = r^B\alpha$, $p = r^R\alpha$. As in the proof of Lemma 3.1, 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 3.1. 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

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

Proof of Proposition 3.7. By substitution, we rewrite the firm's optimization problem (3.11a)-(3.11c):

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

where $\beta(\alpha,\kappa)=\alpha+\frac{c}{\overline{V}\mu(1-\kappa\alpha)}$, which is obtained from subtracting (3.11c) from (3.11b).

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

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 $\kappa=0$,

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

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

At $\kappa=0$, β maximizes (3.9a)-(3.9b). The first-order condition yields:

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

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 (C.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 $\bar{V}$),

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

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

Or equivalently,

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

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

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

Since the left-hand side of (C.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 (C.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. □

C.2 More Details on the Comparison of the Two Referral Programs

C.2.1 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.)

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

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

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

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

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

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

C.2.2 Numerical Studies of Price Adjustments and Throughput Changes

Tables C.1 and C.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 3.6 and 3.7).

Tables C.3 and C.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 throughput may increase or decrease. Combining these two tables with Tables C.1 and C.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 C.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

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

Λ	$V=5$	$V=7.5$	$V=10$	$V=12.5$	$V=15$	$V=17.5$	$V=20$
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

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

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 C.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.

C.3 Observable Queues

Proof of Proposition 3.8. If customers follow threshold strategy n in equilibrium, then for customer seeing queue length $n-1$, it must be rational not to refer; and for customer seeing

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

Λ	$V=5$	$V=7.5$	$V=10$	$V=12.5$	$V=15$	$V=17.5$	$V=20$
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

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

queue length n, it must be rational to refer. This translates to the following conditions:

$$c_r + p\frac{c}{\mu} + (1-p)cW(n;n) > cW(n;n);$$

$$c_r + p\frac{c}{\mu} + (1-p)cW(n+1;n) < cW(n+1;n).$$

Shuffling terms gives

$$W(n;n) < \frac{c_r}{pc} + \frac{1}{\mu} < W(n+1;n).$$

In particular, $n=0$ (everyone refers) if and only if $\frac{c_r}{pc}+\frac{1}{\mu} < W(1;0)$. $n=N-1$ (nobody refers) iff $\frac{c_r}{pc}+\frac{1}{\mu} > W(N-1;N-1) = \frac{N-1}{\mu}$.

Case 1. For $i \le n$, given that all the existing customers in the queue do not make referrals, a successful referral will bring the arriving customer to the head of the queue. Since

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

Λ	$V=5$	$V=7.5$	$V=10$	$V=12.5$	$V=15$	$V=17.5$	$V=20$
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

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

$W(n;n) < \frac{c_r}{pc} + \frac{1}{\mu}$, $W(i;n) < \frac{c_r}{pc} + \frac{1}{\mu}$ for $i \le n$. Therefore,

$$c_r + p\frac{c}{\mu} + (1-p)cW(i;n) > cW(i;n), \quad i \le n.$$

This implies all customers seeing a queue length less than n will not refer.

Case 2. For $i \ge n+1$, making a successful referral will not necessarily move the referring customer to the head of the queue, but will at least move her up by at least n positions. The expected total cost if one refers is bounded from above by

$$c_r + p\frac{c(i-n)}{\mu} + (1-p)cW(i;n), \quad i \ge n+1.$$

Since $W(i;n) - W(i-1;n) \ge \frac{1}{\mu}$,

$$W(i;n) \ge W(n+1;n) + \frac{i-n-1}{\mu} > \frac{c_r}{pc} + \frac{1}{\mu} + \frac{i-n-1}{\mu}, \quad i \ge n+1.$$

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

Λ	$V=5$	$V=7.5$	$V=10$	$V=12.5$	$V=15$	$V=17.5$	$V=20$
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

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

Hence,

$$W(i;n) > \frac{c_r}{pc} + \frac{1}{\mu} + \frac{i-n-1}{\mu}.$$

Shuffling terms yields

$$c_r + p\frac{c(i-n)}{\mu} + (1-p)cW(i;n) < cW(i;n).$$

This implies all customers seeing a queue length at least n will refer. □

Procedure to Compute the Expected Waiting Time Given the threshold strategy n, let $T(i,k)$ be the expected waiting time for a non-referring customer at position i with k customers behind. $i=0,1,\ldots,k=0,1,\ldots$. $T(0,k)=0,\ \forall k$.

Therefore, $W(i;n)=T(i,0)$.

Here is the system of linear equations used to compute $T(i,k)$:

$$\begin{aligned}
T(i,k) &= \frac{1}{\Lambda+\mu} + \frac{\Lambda}{\Lambda+\mu}T(i,k+1) + \frac{\mu}{\Lambda+\mu}T(i-1,k), \quad i+k<n, i\geq 1, k\geq 0,\\
T(i,k) &= \frac{1}{\Lambda+\mu}\\
&\quad + \frac{\Lambda}{\Lambda+\mu}\left[\sum_{g=0}^{N-i-k-2}(1-p)p^g T(g+i,k+1) + p^{N-i-k-1}T(N-k-1,k+1)\right]\\
&\quad + \frac{\mu}{\Lambda+\mu}T(i-1,k), \quad n\leq i+k<N, i\geq 1, k\geq 0,\\
T(i,N-i) &= \frac{1}{\mu} + T(i-1,N-i), \quad i=1,\ldots,N.
\end{aligned}$$

To solve this system of equations, an iterative algorithm can be developed in the same spirit of Hassin and Haviv (1997).

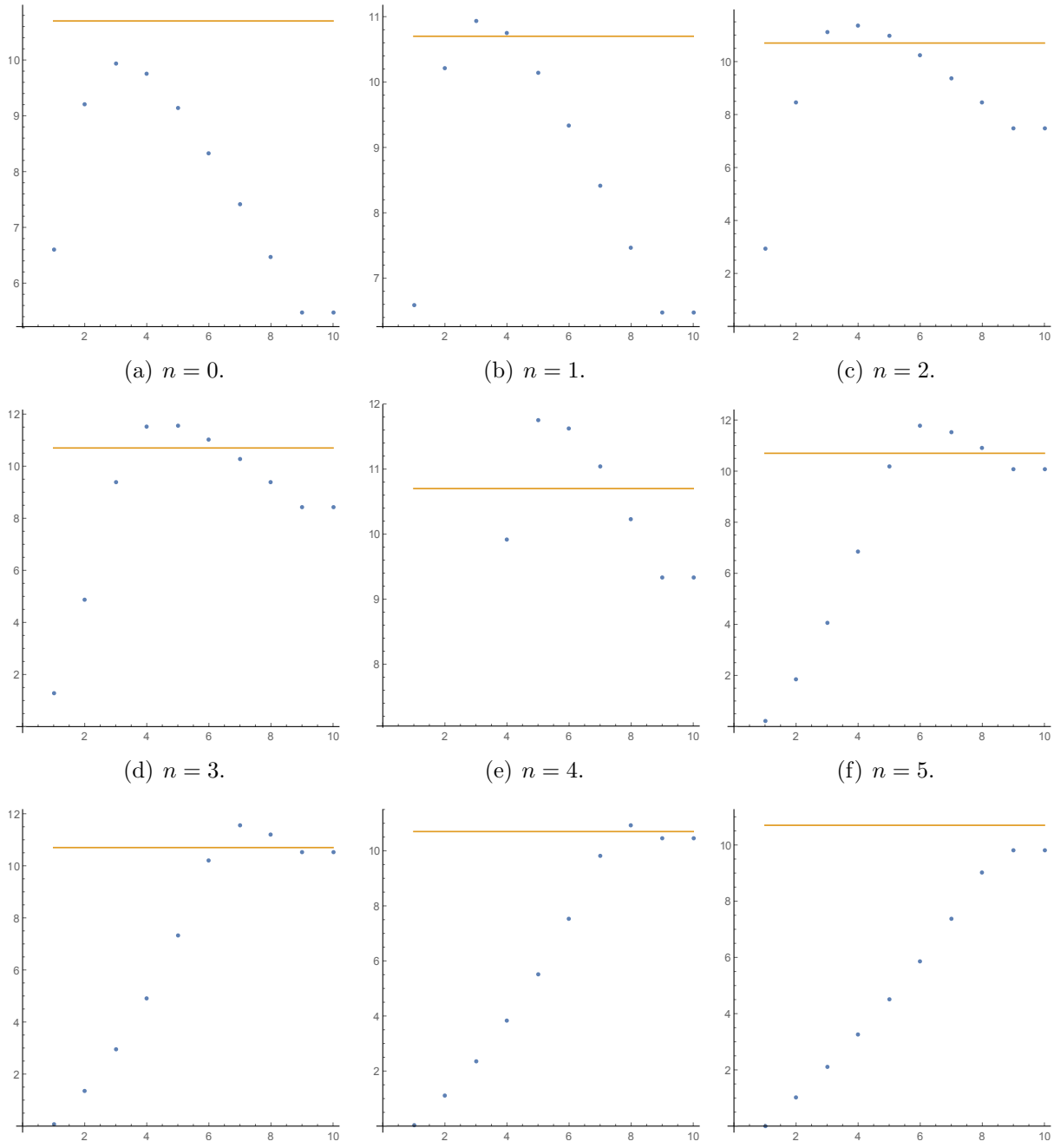
Customers may NOT follow threshold referral strategies. We show an numerical instance in which any pure threshold referral strategy does not constitute an equilibrium. The model primitives are $\mu=1, \Lambda=1.1, p=1, N=10, c=1, c_r=10.7$.

Since $N=10$, possible pure threshold referral strategies are $n=0,1,\ldots,8$. Since $p=1$, referral conversion is guaranteed. Therefore, if customer i makes a referral, her expected waiting time is $\max\{1/\mu,(i-n)/\mu\}$. Thus, threshold strategy n constitutes an equilibrium if and only if

$$\begin{aligned}
c_r &> c\left[W(i;n) - \frac{1}{\mu}\right], \quad i=1,\ldots,n\\
c_r &\leq c\left[W(i;n) - \frac{i-n}{\mu}\right], \quad i=n+1,\ldots,N-2.
\end{aligned}$$

We enumerate all possible $n=0,1,\ldots,8$ and debunk each one of them in Figure C.1.

Figure C.1: Enumeration of pure threshold referral strategies to show none would be sustained in equilibrium.



(a) $n = 0$. (b) $n = 1$. (c) $n = 2$.

(d) $n = 3$. (e) $n = 4$. (f) $n = 5$.

(g) $n = 6$. (h) $n = 7$. (i) $n = 8$.

Note. $\mu = 1, \Lambda = 1.1, p = 1, N = 10, c = 1, c_r = 10.7$. The flat line denotes c_r/c; the dots, $W(i;n) - \max\{1/\mu, (i-n)/\mu\}$ for each i given n. An equilibrium would be sustained if the dots were under the flat line for $i \leq n$ and above for $i > n$.

REFERENCES

Adiri, I., U. Yechiali. 1974. Optimal priority-purchasing and pricing decisions in non-monopoly 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., M. Pavlin. 2016. Optimal price-lead time menus for queues with customer choice: Priorities, pooling and strategic delay. *Management Science* **62**(8) 2412–2436.

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

Allon, G., E. Hanany. 2012. Cutting in line: Social norms in queues. *Management Science* **58**(3) 493–506.

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

Bagnoli, M., T. Bergstrom. 2005. Log-concave probability and its applications. *Economic Theory* **26**(2) 445–469.

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

Barua, B. 2015. Waiting your turn: Wait times for health care in canada, 2015 report. Report, Fraser Institute. Dec. 8, 2015.

Baum, S. 2012. Wait time, bedside manner, knowledge: 5 approaches to doctor rating websites. MedCity News. URL `http://medcitynews.com/2012/08/heres-a-look-at-some-of-the-criteria-5-websites-with-doctor-ratings-use/?rf=1`. August 20, 2012.

BBC News Business. 2014. Mobile firm EE introduces queue jumping charge. BBC. URL `http://www.bbc.com/news/business-28790807`. August 14, 2014.

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

Bulow, J., J. Roberts. 1989. The simple economics of optimal auctions. *Journal of Political Economics* **97**(5) 1060–90.

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.

Chatterjee, K., W. Samuelson. 1983. Bargaining under incomplete information. *Operations Research* **31**(5) 835–851.

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

Chen, H., Q. Qian, A. Zhang. 2015. Would allowing privately funded health care reduce public waiting time? theory and empirical evidence from canadian joint replacement surgery data. *Production and Operations Management* **24**(4) 605–618.

Chen, Y., J. Meinecke, P. Sivey. 2016. A theory of waiting time reporting and quality signaling. *Health Economics* **25**(11) 1355–1371.

Coase, R. 1960. The problem of social cost. *Journal of Law and Economics* **3**(1) 1–44.

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

Cox, D.R., W. Smith. 1961. *Queues*. Methuen, London.

Cramton, P., R. Gibbons, P. Klemperer. 1987. Dissolving a partnership efficiently. *Econometrica* **55**(3) 615–632.

Cui, S., X. Su, S. Veeraraghavan. 2016. A model of rational retrials in queues. Working paper, Wharton School, University of Pennsylvania.

Cui, S., J. Wang, Z. Wang. 2017. Equilibrium strategies in $M/M/1$ queues with priorities. Working paper, Georgetown University.

Davidson, C. 1988. Equilibrium in servicing industries: An economic application of queuing theory. *The Journal of Business* **61**(3) pp. 347–367.

DeAmicis, Carmel. 2014. New dining app reserve hopes to challenge opentable by charging customers instead of restaurants. GigaOm. URL `https://gigaom.com/2014/10/28/new-dining-app-reserve-hopes-to-challenge-opentable-by-charging-customers-instead-of-restaurants/`. Oct. 28, 2014.

Debo, L., C. Parlour, U. Rajan. 2012. Signaling quality via queues. *Management Science* **58**(5) 876–891.

Debo, L., S. Veeraraghavan. 2014. Equilibrium in queues under unknown service times and service value. *Operations Research* **62**(1) 38–57.

Dranove, D., D. Kessler, M. Mcclellan, M. Satterthwaite. 2003. Is more information better? the effects of "report cards" on health care providers. *Journal of Political Economy* **111**(3) 555–588.

Dranove, D., M. Satterthwaite. 1992. Monopolistic competition when price and quality are imperfectly observable. *RAND Journal of Economics* **23**(4) 518–534.

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

El Hajia, A., S. Onderstal. 2015. Trading places: An experimental comparison of reallocation mechanisms for priority queuing. Working paper, University of Amsterdam.

Fotaki, M., M. Roland, A. Boyd, R. McDonald, R. Scheaff, L. Smith. 2008. What benefits will choice bring to patients? literature review and assessment of implications. *J Health Serv Res Policy* **13**(3) 178–84.

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.

Gershkov, A., P. Schweinzer. 2010. When queueing is better than push and shove. *International Journal of Game Theory* **39**(3) 409–430.

Glazer, A., R. Hassin. 1983. Search among queues. Unpublished report, Tel Aviv University.

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.

Graham, C. 2000. Chaoticity on path space for a queueing network with selection of the shortest queue among several. *Journal of Applied Probability* **37**(1) 198–211.

Gray, K. 2009. Property in a queue. G. S. Alexander, E. M. Penalver, eds., *Property and Community*. Oxford University Press, New York, 165–195.

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. 1996. On the advantage of being the first server. *Management Science* **42**(4) 618–623.

Hassin, R., M. Haviv. 1994. Equilibrium strategies and the value of informaiton in a two line queueing system with threshold jockeying. *Communications in Statistics. Stochastic Models* **10**(2) 415–435.

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*. Springer Science & Business Media, New York.

Hassin, R., R. Roet-Green. 2015. The impact of inspection costs on equilibrium in a queueing system with parallel servers. Working paper, Tel Aviv University.

Hassin, R., R. Roet-Green. 2016. The impact of inspection cost on equilibrium, revenue, and social-welfare in a single server queue. *Operations Research* Forthcoming.

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

Honka, E., P. Chintagunta. 2017. Simultaneous or sequential? search strategies in the U.S. auto insurance industry. *Marketing Science* **36**(1) 21–42.

Hu, M., Y. Li, J. Wang. 2016. Efficient ignorance: Information heterogeneity in a queue. *Management Science* Forthcoming.

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

Kahneman, D., J. L. Knetsch, R. H. Thaler. 1990. Experimental tests of the endowment effect and the Coase theorem. *Journal of Political Economy* **98**(6) 1325–1348.

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

Kim, J., P. Albuquerque, B. Bronnenberg. 2010. Online demand under limited consumer search. *Marketing Science* **29**(6) 1001–1023.

Kirkey, S. 2014. Canadian physicians get advice on how to handle 'rate my doctor' websites. Canada.com. URL `http://o.canada.com/news/canadian-physicians-get-advice-on-how-to-handle-rate-my-doctor-websites`. Sep 21, 2014.

Kittsteiner, T. 2003. Partnerships and double auctions with interdependent valuations. *Games and Economic Behavior* **44**(1) 54–76.

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.

Kreindler, Sara A. 2010. Policy strategies to reduce waits for elective care: a synthesis of international evidence. *British Medical Bulletin* **95** 7–32.

Leclerc, F., B. H. Schmitt, L. Dubé. 1995. Waiting time and decision making: Is time like money? *Journal of Consumer Research* **22**(1) 110–119.

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.

Luczak, M. J., C. McDiarmid. 2006. On the maximum queue length in the supermarket model. *Ann. Probab.* **34**(2) 493–527.

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

Mann, L. 1969. Queue culture: the waiting time line as a social system. *American Journal of Sociology* **75**(3) 340–354.

McCall, B.P., J.J. McCall. 2008. *The Economics of Search*, vol. 1. Routledge, New York.

McCall, J. J. 1970. Economics of information and job search. *Quarterly Journal of Economics* **84**(1) 113–126.

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

Miller, D. R. 1981. Computation of steady-state probabilities for M/M/1 priority queues. *Operations Research* **29**(5).

Mitzenmacher, M. 2001. The power of two choices in randomized load balancing. *IEEE Trans. Parallel Distrib. Syst.* **12**(10) 1094–1104.

Morgan, P., R. Manning. 1985. Optimal search. *Econometrica* **53**(4) 923–944.

Mortensen, D. 1970. Job search, the duration of unemployment, and the Phillips curve. *American Economic Review* **60**(5) 847–62.

Myerson, R. B. 1981. Optimal auction design. *Mathematics of Operations Research* **6**(1) 58–73.

Myerson, R. B., M. A. Satterthwaite. 1983. Efficient mechanisms for bilateral trading. *Journal of Economic Theory* **29**(2) 265–281.

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

Nelson, P. 1970. Information and consumer behavior. *Journal of Political Economy* **78**(2) 311–329.

Oberholzer-Gee, F. 2006. A market for time fairness and efficiency in waiting lines. *Kyklos* **59**(3) 427–440.

Rachlis, Michael M. 2005. Public solutions to health care wait lists. Report, Canadian Center for Policy Alternatives.

Rayo, L. 2013. Monopolistic signal provision. *The B.E. Journal of Theoretical Economics,* **13**(1) 27–58.

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/`.

Rochet, J., P. Choné. 1998. Ironing, sweeping, and multidimensional screening. *Econometrica* **66**(4) 783–826.

Rogerson, R., R. Shimer, R. Wright. 2005. Search-theoretic models of the labor market: A survey. *Journal of Economic Literature* **43**(4) 959–988.

Rosenblum, D. M. 1992. Allocation of waiting time by trading in position on a $G/M/S$ queue. *Operations Research* **40** S338–S342.

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.

Satterthwaite, M. A., S. R. Williams. 1989. Bilateral trade with the sealed bid k-double auction: Existence and efficiency. *Journal of Economic Theory* **48**(1) 107–133.

Sattinger, M. 2010. Queueing and searching. Working paper, University at Albany, SUNY - Department of Economics.

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

Schwartz, B. 1975. *Queueing and Waiting*. University of Chicago Press, Chicago.

Shaked, M., J.G. Shanthikumar. 2007. *Stochastic Orders*. Springer Series in Statistics, Springer.

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`.

Sicilani, L., J. Hurst. 2004. Explaining waiting-time variations for elective surgery across OECD countries. *OECD Economic Studies* **2004**(1) 95–123.

Siciliani, L., J. Hurst. 2005. Tackling excessive waiting times for elective surgery: a comparative analysis of policies in 12 OECD countries. *Health Policy* **72**(2) 201–215.

Siddique, H. 2015. NHS waiting times for elective surgery spiked last year, report finds. The Guardian. URL `https://www.theguardian.com/society/2015/jul/24/nhs-waiting-times-elective-surgery-spike-patients-association-report`. July 24, 2015.

Stigler, G. J. 1961. The economics of information. *Journal of Political Economy* **69**(3) 213–225.

Stross, Randall. 2010. The online reservations that restaurants love to hate. The New York Times. URL `http://www.nytimes.com/2010/12/12/business/12digi.html`. Dec. 11, 2010.

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.

Thomson, S., A. Dixon. 2006. Choices in health care: the European experience. *J Health Serv Res Policy* **11**(3) 167–71.

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.

Turner, S. 1998. The effect of increasing routing choice on resource pooling. *Probability in the Engineering and Informational Sciences* **12** 109–123.

Ursu, R. 2016. The power of rankings: quantifying the effect of rankings on online consumer search and purchase decisions. Working paper, New York University.

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

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.

Vvedenskaya, N., R. Dobrushin, F. Karpelevich. 1996. Queueing system with selection of the shortest of two queues: An asymptotic approach. *Problems of Information Transmission* **32** 15–27.

Weitzman, M. 1979. Optimal search for the best alternative. *Econometrica* **47**(3) 641–54.

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.

WTA report card. 2015. Eliminating code gridlock in canadas health care system. Report, Waiting Time Alliance.

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.

Xu, J., B. Hajek. 2013. The supermarket game. *Stochastic Systems* **3**(2) 405–441.

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.