Title
Essays on Service Operations Systems: Incentives, Information Asymmetries and Bounded Rationalities

Author
He, Qiaochu

Publication Date
2016

Peer reviewed|Thesis/dissertation

Essays on Service Operations Systems: Incentives, Information Asymmetries and Bounded Rationalities

by

Qiao-Chu He

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering - Industrial Engineering and Operations Research

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Zuo-Jun Shen, Chair, Chair
Professor Philip M. Kaminsky
Associate Professor Terry A. Taylor

Spring 2016

Essays on Service Operations Systems: Incentives, Information Asymmetries and Bounded Rationalities

Abstract

Essays on Service Operations Systems: Incentives, Information Asymmetries and Bounded Rationalities

by

Qiao-Chu He

Doctor of Philosophy in Engineering - Industrial Engineering and Operations Research

University of California, Berkeley

Professor Zuo-Jun Shen, Chair, Chair

This dissertation is concerned with service operations systems with considerations of incentives, information asymmetries and bounded rationalities. Chapter 1 provides an overview of the dissertation.

In Chapter 2, motivated by the information service operations for the agricultural sectors in the developing economies, we propose a Cournot quantity competition model with price uncertainty, wherein the marketing boards of farmers' cooperatives have the options to obtain costly private information, and form information sharing coalitions. We study the social value of market information and the incentives for information sharing among farmers.

In Chapter 3, we offer a behavioral (bounded rationality) theory to explain product/technology adoption puzzle: Why superior investment goods are not widely purchased by consumers? We show that present-bias encourages procrastination, but discourages strategic consumer behavior. Advance selling is beneficial not only to the consumers as a commitment device, but also to the seller as a price discrimination instrument.

In Chapter 4, motivated by the fresh-product delivery industry, we propose a model of service operations systems in which customers are heterogeneous both in terms of their private delay sensitivity and taste preference. The service provider maximizes revenue through jointly optimal pricing strategies, steady-state scheduling rules, and probabilistic routing policies under information asymmetry. Our results guide service mechanism design using substitution strategies.

In Chapter 5, motivated by the puzzle of excessively long queue for low quality service in tourism and healthcare industries, we study the customers learning behaviors in the service operations systems, when they hold incorrect beliefs about the population distribution. We highlight a simple behavioral explanation for the blind "buying frenzy" in service systems with low quality: The customers under-estimate others' patience and are trapped in a false optimism about the service quality.

Chapter 6 concludes the dissertation with a summary of the main results and policy recommendations.

To my parents.

Contents

List of Figures

List of Tables

Acknowledgments

First and foremost, I would like to express gratitude for my adviser, Professor Ying-Ju Chen. He always tolerates my stupidity and naivety, and never ceases to believe in me. Without his guidance and encouragement, I would never be able to finish this dissertation. I would also like to thank my adviser Professor Zuo-Jun Max Shen. His advice on research and academic career path is most valuable to me. I would like to thank everyone in his research group, since I benefited greatly through group meetings and discussions.

I would also like to thank Professor Rhonda Righter, for her good advice on course learning, teaching and research. It was a privilege and honor for me to work with her. I have also benefited from valuable comments and feedback from my dissertation committee members, Professor Philip Kaminsky and Professor Terry Taylor, for which I am truly grateful.

Last, but by no means least, I would like to thank all my friends at Berkeley. Without their company, this journey would also be impossible.

Chapter 1

Introduction

This dissertation consists of four essays on service systems, motivated by information service operations in agriculture, distribution of product/technology in developing economies, fresh-product delivery services, and tourism industries. Conflicting objectives are the basic ingredients in these four separate yet related theories: Revenue-maximizing farmers receive inferior social welfare, myopic consumers do not adopt investment goods, and a queue-joining customer *causes* congestion to others. This paradigm for the analysis of service systems (i.e., by considering rational/self-interested agents) is further enriched by decentralized information and learning: In agriculture, governments provide market forecast as guidance to farmers' production planning. In service systems, "long queue signals good service". Finally, we deviate from this paradigm of analysis by considering bounded rationalities: In developing economies, *present-bias* leads to lack of self-control, which has causal relationship with poverty. In service systems, I show that *projection effect* can easily explain "long queue for bad service". These deviations are made carefully, and only when they are necessary. More detailed descriptions are presented in the following sections.

1.1 On the Formation of Farmer Producer Organizations

In Chapter 2, we study the incentives for farmers' cooperatives in developing economies to conglomerate and form farmer producer organizations (FPOs). We focus on the FPOs' efforts in linking farmers by integrating market information. We propose a Cournot quantity competition model with price uncertainty, wherein the marketing boards of farmers' cooperatives have the options to obtain costly private information, and form information sharing coalitions. Through a social responsibility lens, we examine how the information service operations for non-governmental organizations (NGOs) can improve the farmers' welfare and reduce poverty.

We find that when there are only two farmers' cooperatives, they have no incentive to share market information in equilibrium. In general, neither a single coalition nor social

isolation is sustained in equilibrium. Multiple competing FPOs are formed when the public information provision is low, and the marketing boards *obtain too much private information* while *use it too little*. In this case, the farmers' revenues decrease in the precision of public information provided by the NGOs. On the other hand, when the public information provision is high, our model predicts a single dominating FPO, while all the non-affiliated farmers' cooperatives do not share information. In this case, the farmers' revenues increase in the public information provision. From the policy perspective, our model offers insights on the NGO's dual roles in providing market information as well as mobilizing farmers to build FPOs.

1.2 Selling Investment Goods with Present-Biased Consumers

In Chapter 3, we propose a stylized monopoly pricing model with investment goods, wherein consumers suffer from *present-bias*: Consumers procrastinate purchase decisions but make no purchase later due to lack of self-control. This bounded rationality is used to explain why certain superior investment goods are not widely adopted in developing economies.

We show that advance selling can be beneficial both to the seller as an inter-temporal discrimination instrument, and to the consumers as a commitment device. Present-bias can either increase or decrease aggregate product adoption, as it encourages procrastination behaviors while discourages strategic consumer behaviors. When a donor desires to stimulate the product adoption by subsidizing the consumers, we recommend timely subsidy in the advance-market to disincentivize delay in purchase. Surprisingly, increasing public awareness of lack of self-control may or may not help in general, as it can either increase or decrease the donor's equilibrium subsidy level in the spot-market.

1.3 Revenue-maximizing Pricing and Scheduling Strategies in Service Systems with Horizontal Substitutions

In Chapter 4, we propose a model of service operations systems in which customers are heterogeneous both in terms of their private delay sensitivity and taste preference. The service provider maximizes revenue through jointly optimal pricing strategies, steady-state scheduling rules, and probabilistic routing policies under information asymmetry. The impact of *horizontal substitutions* is twofold: It provides instrument to balance the traffic intensities between horizontal differentiated services, however, the service provider should sacrifice information rent to create incentives for customers to truthfully report their taste preference. If only the taste attribute is observable, the service provider still needs to pay information rent to patient customers. If the delay sensitivity is observable, the complete information bench-

mark could be restored. We compare this basic model with an alternative *server-specific* mechanism, in which the service provider does not differentiate service with respect to customers' taste preference *ex post*. The optimal scheduling policies in equilibrium combine the *absolute preemptive priority* and *strategically inserted delay*, and we identify new roles of these two non-trivial queueing disciplines in the flexible system context. In particular, when one queue accommodates a large population of impatient customers, it may be desirable to strategically idle the server in the other queue. This phenomenon is new to the literature as the existing papers focus exclusively on a single-server system wherein strategic delays take place *within* the same queue. Finally, we show that the revenue gap between the basic model and the server-specific model is small, while the latter service mechanism is easier to implement.

1.4 Learning with Projection Effects in Service Operations Systems

In Chapter 5, we study the customers' learning behaviors in the service operations systems, when customers hold incorrect beliefs about the population distribution. We propose a single-server queueing model with observable queue length, in which the customers are heterogeneous both in terms of their delay sensitivity and information precision about the unknown service quality. We compare the system performances when the customers suffer from the *(reversed-) projection effects*, i.e., bounded rationalities under which the customers expect the others to be more (less) similar to themselves than reality, in terms of their delay sensitivity.

Ironically, under projection bias, customers who are more averse to waiting will react more sensitively to the observed long queue, which leads to over-estimation of the service quality and waiting on the long queue. Such bounded rationalities impede effective learning by inducing decision errors, which could reduce the social welfare due to blind "buying frenzy" even if the service quality is low. Finally, the queue lengths are the longest when the impatient customers suffer from the projection bias while the patient customers suffer from the reversed-projection bias, because all the uninformed customers are simultaneously trapped in the false optimism situations by under-estimating others' patience.

Chapter 2

On the Formation of Farmer Producer Organizations

2.1 Introduction

Background and motivation

The year 2014 is being observed as the "Year of Farmer Producer Organizations (FPOs)" by the Government of India. The concept of FPOs was introduced by the Indian government in 2002, which attempts to establish basic business principles within farming communities, to bring industry and agriculture closer together, and to boost rural development (Trebbin and Hassler, 2012). Despite the multi-faceted role of FPOs, we focus on their efforts in linking farmers by integrating market information, with the objective to improve farmers' welfare. Examples of FPOs are well documented in many other developing economies as well, e.g., South Africa (Nieuwoudt, 1987), China (Jia and Huang, 2011), Vietnam (Moustier et al., 2010), Honduras and El Salvador (Hellin et al., 2009), Peru and Ecuador (Devaux et al., 2009), Kenya (Fischer and Qaim, 2012) and Uganda (Kaganzi et al., 2009).

Non-governmental organizations (NGOs) play an important role in linking farmers to markets. In particular, we have observed the shift from production intervention towards information intervention. Literature indicates that the effects of production intervention are not always positive. For example, Michelson et al. (2012) report that United States Agency for International Development funded four NGOs in Nicaragua, who work with three farmers' cooperatives to sign a three-year production contract with Walmart. However, the prices paid by Walmart are significantly lower than the traditional market. Alternatively, NGOs start to connect local farmers' cooperatives as *information sharing coalitions*, by integrating market information but not dictating the production decisions for members[1]. As a result, we observe a hierarchical structure shown in Figure 2.1: At local level, the building blocks of FPOs are

[1]The readers are referred to the "Policy and Process Guidelines for Farmer Producer Organization" by India government for detailed descriptions on the operational principles of FPOs.

marketing boards, or marketing cooperatives which represent groups of farmers[2]. A local farmers' cooperative may impose quota on each farmer's production quantity by formal or informal contracts (Nieuwoudt, 1987), upon which the FPO links those cooperatives together and with the market[3].

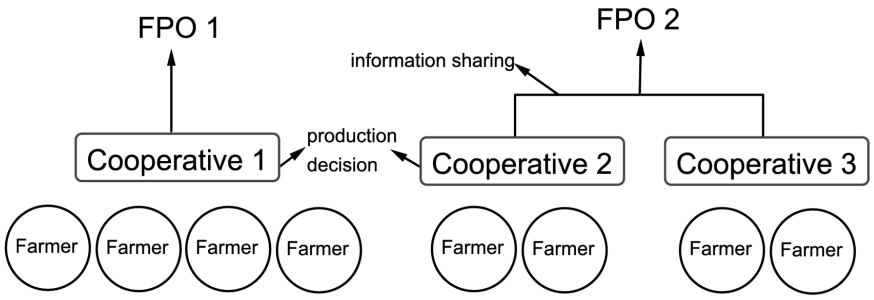

Figure 2.1: Example of the hierarchical structure of FPOs.

For example, *Vrutti* is a *not-for-profit* organization who mobilizes farmers in groups at village level to build FPOs [4]. Vrutti is currently working with 4000 soy farmers in Madhya Pradesh, and they establish a peer learning system among lead farmers, each representing a farmers' cooperative at community level, on top of which an FPO is shaped as a legal form of this network of farmers' cooperatives. In this case, Vrutti creates Agriculture Enterprise Facilitation Centre as an information exchange platform to incubate FPOs. In another example, Qiao and Yu (2013) document how a local Chinese watermelon farmers' cooperative expands to an FPO, serving more than 50 fruit wholesale markets in more than 20 cities. Qiao and Yu (2013) ascribe their success (partially) to "information spillover" of the kind which we are addressing to. They also observe the asymmetric information structure: The organization often sends a representative to government meetings and seminars who brings back to the farmers useful market information, but non-members cannot avail of this information. Devaux et al. (2009) document the case of *Iniciativa Papa Andina*, a partnership program linking potatoes farmers to market in Bolivia, Ecuador and Peru. In Ecuado, for example, 24 farmers' groups are created based on which a national organization, the Consortium of Small Potato Producers, is established to support joint marketing activities, including "knowledge sharing and social learning". Fischer and Qaim (2012) show that mobile phone ownership is an important determinant of FPO membership in Kenya. This

[2]We address the information asymmetry among farmers' cooperatives rather than individual farmers. This is because much of the information is already being shared at village level, and smallholdings from the same village typically grows the same crops.

[3]See (http://www.vrutti.org/index.php/projects/aefc) for further illustration.

[4]Funded by the Indian government and the World Bank, Vrutti has been working with over 45 FPOs, each consisting of hundreds to thousands of farmers. See (http://www.vrutti.org/index.php/projects/fpo) for a list of FPOs it support.

serves as empirical evidence of the FPO's role concerning information exchange in developing economies.

As we have observed, NGOs (such as Vrutti) are heavily involved with the information service operations for the agricultural sector in the developing economies, by providing leadership and operational support for the formation and sustainable development of FPOs. We shall examine those NGOs through a *social responsibility lens*, and explore the functionality of FPOs as information sharing coalitions among farmers' cooperatives.

Research questions and modeling framework

In this chapter, we shall address the following two central research questions:

1. What are the incentives for farmers' cooperatives to conglomerate and form FPOs?

2. Does market information provision improve farmers' welfare?

The first research question is partially answered by An et al. (2015), by focusing on the physical aspects such as production cost reduction, intermediaries elimination etc. In contrast, we examine the informational aspects in this chapter. Our research also differs by endogenizing the organization formation process. The second question is addressed by Chen and Tang (2015), who show that private information does create value for the farmers while the public signal might not. However, the two questions are not independent: When information sharing coalitions are formed, does additional market information still create value for the farmer? Conversely, when farmers are provided with direct information acquisition channels (either private or public), will the indirect channel (information sharing coalitions) still be useful? By carefully analyzing those interplays between the two issues, we contribute to a holistic view of the information service operations for agriculture in the developing economies.

We propose a stylized model in which farmers' cooperatives obtain market price information from direct information acquisition, information sharing and free public sources. The marketing boards of those cooperatives make production decisions based on the market information available to them. There is a common market for the farmers to sell their products to the buyers. The farmers' cooperatives engage in Cournot competition under uncertain market prices.

The Cournot competition framework among farmers' cooperatives follows recent stream of literature, e.g., An et al. (2015), Chen and Tang (2015), etc. We justify this setup from the following aspects. Firstly, we interpret "farmers" as cooperatives. As articulated in An et al. (2015), many smallholder farmers organize themselves (or through external help) as cooperatives in developing countries such as India. These cooperatives may impose production quotas (Nieuwoudt, 1987), which serves as a justification for quantity competition. In Europe, dairy produce quota is used by the European Economic Community to control the market price of dairy products. While individual farmers may be ignorant about their

influence and the market interactions, farm cooperatives do take competition and other factors into consideration when making production decisions. Secondly, Cournot competition is supported by evidence in agricultural setting. The existence of quantity competition in various agricultural product markets in general is supported by some well-known empirical studies, e.g., Deodhar and Sheldon (1996). The empirical evidence suggests that Cournot competition is appropriate, at least, for some agricultural markets. Specifically, Cournot competition is commonly used for modeling market competition of commodities for crops that take substantial time to produce so that the output cannot be adjusted quickly (Brander and Spencer, 1985). Carter and MacLaren (1994) also indicate that in these situations, Cournot competition is a good approximation of the real economic decision-making. Furthermore, Moustier et al. (2010) document Vietnamese rice FPOs, which show that it is possible for the farmers' organization to secure a significance presence in the local supply chain in developing economies. Finally, using a unique transaction-level data from Ethiopia agricultural market, Osborne (2005) empirically tests the existence of quantity competition (although the competition is not perfect).

To make wise production decisions for farmers, it is crucial for the marketing boards of the farmers' cooperatives to acquire accurate market information, e.g., price forecasts. It should be noted that we focus exclusively on the long-term market information. For example, on the website of Government of India (www.agmarknet.nic.in), there are available reports and expert analysis on long-term "price trends" and "price behavior", which serve as guidelines for production planning.

We incorporate both public and private information channels in the model. Examples of private market information providers include Reuters Market Light (RML), which tracks the prices of 50 commodities over 1000 markets, and provides a short messaging service costing 60 rupees ($1.50) a month in subscription for Indian farmers. Examples of public market information providers include Indian Tobacco Company (ITC). ITC also establishes its e-Choupal network which covers numerous villages through the Internet-based kiosks for free, which helps farmers to obtain higher selling price for their soybeans in Central India (Goyal, 2011). In Kenya and Mali, an NGO launched a weekly hour-long radio program called Mali Shambani which discusses market price trends. The India Ministry of Agriculture launched the Kisan Call Centers in 2004 to deliver free information services to farmers over the phone.

To model the endogenous formation of FPOs as information sharing coalitions, we let the marketing boards of farmers' cooperatives decide whether to build communication channel with each other. We assume mutual information exchange: Once a connection is made, the two connected parties observe each other's private information. As we are concerned with the long-term interactions among the farmers, we abstract away from the iterative process of network evolution and information updating. As a result, the long-run information structures are homogeneous within a coalition but heterogeneous across its disconnected counterparts. Eventually, the emerging coalitions are interpreted as FPOs, which integrate market information within but are separated among each other.

Summary of results

We first present the results of a model with two farmers (from now on, we use the term "farmer" to indicate the smallest entity who makes the production quantity decisions, e.g., the marketing boards of farmers' cooperatives or relatively "big" individual farmers). In particular, we find that no information sharing coalition is formed in any Nash equilibrium. Intuitively, the information asymmetry due to isolation is beneficial to the farmers since it alleviates over-reaction to the common signals, and relieves the quantity competition stress. When both disconnected farmers obtain private signals, they acquire excessive information which leads to suboptimal revenues. Furthermore, the farmers' revenues decrease in the public information provision, since they *obtain too much private information* and *use it too little*.

When there are multiple farmers who form a single information sharing coalition, their aggregate payoff is independent of the public information provided by the NGOs (or government). However, this coalition will collapse in the context which we are addressing to in this chapter, i.e., when the market uncertainty is still high with the available public information. At the other extreme, social isolation is also not sustained in equilibrium when the population is large. In general, some degree of information sharing will take place in equilibrium, and we should expect to observe multiple FPOs who integrate market information within but are isolated organization-wise.

When the public information provision is low, each individual farmer's revenue is quasi-concave in the private information provision of her FPO. This is driven by the interaction of the *competition effect* and the *congestion effect*, as we shall elaborate in the analysis. In this case, the farmers' aggregate payoff decreases in the public information provision. On the other hand, when the public information provision is high, each individual farmer either obtains as much private information as possible, or no information at all. This extreme information acquisition strategy leads to a single dominating FPO, while all the non-affiliated farmers are isolated. We refer to this as the *polarization effect*, under which a fair allocation of revenue is achieved among farmers, and their aggregate revenue increases in the public information provision.

The rest of this chapter is organized as follows. Section 2.2 reviews relevant literature. Section 2.3 introduces our model setup. In Section 2.4, we carry out the analysis. Section 2.5 provides extensions. Section 2.6 concludes. Major proofs are provided in the appendix.

2.2 Literature Review

Our work falls into the rising research agenda on supply chain management in developing economies, initiated by Sodhi and Tang (2014). Chen et al. (2013) examine the ITC e-Choupal network and discuss how it substantially changes the information and material flows. Chen et al. (2014) further study the peer-to-peer information sharing in Avaaj Otalo. They show that the responses of the knowledgeable farmers are always less informative than those

of the experts. Tang et al. (2014) focus on the issue of whether each farmer should *utilize* market information when both market demand and process yield are uncertain. An et al. (2015) study the impacts of aggregating farmers through formal or informal cooperatives, which include: (1) reducing production cost; (2) increasing/stabilizing process yield; (3) increasing brand awareness; (4) eliminating unnecessary intermediaries; and (5) eliminating price uncertainty. They show that it is beneficial for a farmer to join the aggregation only when the size of the aggregation is below a certain threshold. While they consider a single exogenous aggregation, we endogenize the coalition formation process. The readers are referred to Sodhi and Tang (2014) for a more complete literature survey and future research directions. Our model contributes to this exciting stream of literature in that we endogenize the private information acquisition, and further illustrate the farmers' incentives to form information sharing coalitions.

Our stylized Oligopoly competition model under incomplete information has long roots in economics. Our work is closest to the models on the endogenous information sharing networks in oligopoly. In the earlier economics literature, it is a pervasive view that information sharing equilibria is not sustainable for the Cournot competition with common demand uncertainty, e.g., Gal-Or (1985) and Vives (1984). A general framework is proposed by Raith (1996), in which those earlier results are summarized. More recently, Currarini and Feri (2014) and Lee (2014) analyze models similar to ours. Due to exogenous information provision and symmetry assumption in their models, an empty network is the unique equilibrium outcome by allowing multi-player deviation. The insight behind those pessimistic predications is that, a firm with exogenous private knowledge about the common market uncertainty can enjoy monopoly rent on information, so that it will not reveal such knowledge to an rival. We differ fundamentally from those existing results by predicting asymmetric partial information sharing among oligopolies. Similar results are observed in the literature of supply chain horizontal information sharing. We highlight two papers on demand forecasting and tacit cartel formation, respectively. Shin and Tunca (2010) show that downstream firms under Cournot competition over-invest in demand forecasting. Li (2002) proposes a two-level supply chain model with multiple competing downstream firms, and they show that downstream firms refuse to share demand information.

Motivated by the agricultural production in developing economies, our model incorporates several interesting features worthy of literature review. (1) Endogenous private information acquisition. Vives (1988) shows that the investments in information are strategic substitutes in Cournot competition, and Hellwig and Veldkamp (2009) generalize the discussion. (2) The value of public information provision. Morris and Shin (2002) are among the first to show the adverse effect of public information in a coordination game with strategic externality. Colombo et al. (2014) find that public information crowds out the private information acquisition and thus reduces social welfare in certain regimes. (3) Modeling information diffusion in social network. We model the long-term repetitive interactions such that agents can observe their indirect neighbors' signals (global observability), rather than one-time interactions such that agents can only observe their immediate neighbors' signals (local observability). The reality should be somewhere in between, due to communication

frictions (Myatt and Wallace, 2012), limited memory or sampling costs (Çelen and Kariv, 2004), and delays (Acemoglu et al., 2014). (4) Endogenous network formation. The micro-structure in our model does not fit under typical assumptions in literature, i.e., locally additive information externality, and the concavity of the payoff function with respect to the amount of information (Galeotti and Goyal, 2010). Thus, it is interesting to compare our predictions with similar models, e.g., "star" network (Bloch and Dutta, 2009), "core-peripheral" network (Koenig, 2012), "information ring" (Acemoglu et al., 2014), and observe how different preference assumptions for information drive the difference in predictions. The readers are referred to Jackson (2010) for an excellent review of a broader literature in social and economics networks, and the formation of non-informational networks as well.

2.3 Model

Consider n farmers (she) who produce and sell substitutable agricultural products through a common market[5]. The farmers, indexed by the set $N = \{1, 2, \cdots, n\}$, are homogeneous *ex ante* and engage in a Cournot competition[6]. Suppose that farmer i produces q_i units of products at a cost $c \cdot q_i$, and the aggregate production quantity is denoted as $Q = \sum_{i=1}^{i=n} q_i$. As we abstract away from the demand side structure, we assume that the actual market clearing price $P(Q)$ is linearly decreasing in Q, i.e., $P(Q) = a - bQ + u$, where $u = N(0, \alpha^{-1})$ captures the price uncertainty. The parameter α is the information precision *a priori*, which is public. To deliver a clear presentation of major results, we assume that the constant a is large enough, such that the market clearing price is non-negative with a high probability[7].

Information acquisition. Farmer i has the option to obtain a private signal with precision γ_i at a cost of $r > 0$. The level of private information provision γ_i depends on her search efforts. To convey the major messages, the acquisition cost is represented by an indicator function, i.e., $r \cdot \delta\{\gamma_i > 0\}$. While we assume that the cost r is a fixed constant, regardless of the signal precision, the model can be extended to incorporate general cost structure. The cost of information acquisition is interpreted as the marketing boards paying for market research, Internet access, etc. Consequently, the farmers can observe signals concerning the market uncertainty, i.e., $x_i = u + \epsilon_i$, where $\epsilon_i \sim N(0, \gamma_i^{-1})$, for $\forall i \in N$. The realizations of the signals are private, while their precisions $\{\gamma_i\}$'s are common knowledge. We assume that all the signals obtained by direct information acquisition are pairwise independent. In addition, all the farmers also receive a public signal $x_0 = u + \epsilon_0$, where $\epsilon_0 \sim N(0, \beta^{-1})$, which captures the public market information provided by the NGOs (or the government) free of charge. We restrict ourselves to the *information-scarce regime*

[5]Recall that we interpret "farmers" as the smallest entities who make the production quantity decisions, e.g., the marketing boards of small farmers' cooperatives or relatively "big" individual farmers.

[6]Qiao and Yu (2013) document that farmers' organization tends to absorb homogeneous members, because this practice not only enhances the training effect, but also protects the interests of the old members. We also discuss the heterogeneous case in Section 2.5.

[7]This assumption allows a good approximation by ignoring the kink on the boundary, which is common in the literature, e.g., Li and Zhang (2008). We follow this precedent of sidestepping the issue.

where $\frac{\beta}{\alpha}$ is small, i.e., market uncertainty is high compared with the public information provision. This turns out to be a more interesting and realistic regime, while otherwise the farmers take no initiative to search for information due to the saturation of public information.

FPOs. The formation process of FPOs starts with the establishment of information exchange network. Suppose that farmer i can form a link with farmer j at a cost $k > 0$. The linkage costs can be interpreted as hassle costs in network establishment, which include social mobilization, transportation costs, and negotiations to reach information sharing agreements. Alternatively, the long-run apportionments of those costs can be interpreted as the membership fees for joining the FPO.[8]

Our key assumption about the communication protocol is that the information exchange is *bilateral*, *truthful* and *exclusive*; this means that farmer i is not allowed to observe farmer j's signal unless it reveals its own signal to firm j, and the signals are truthfully revealed. Such protocols are common in the literature, e.g., Currarini and Feri (2014) and Lee (2014). We justify the protocol by the verifiability of field growth, as well as the observation that the formation of FPOs is a long-term process involving farmers' repetitive interactions. As it will soon be explained in detail, this setup leads to complete information sharing within an FPO. Thus, an FPO serves as an information sharing coalition, institutionalizing the information sharing within the organization and the isolation across farmers' cooperatives.

To be specific, we use $\mathbf{g}_i = (g_{i1}, g_{i2}, \cdots, g_{in})$ to denote the linkage decisions, where $g_{ij} = \{0, 1\}$, for $\forall j \in N \setminus i$. For consistency, we assume that $g_{ii} = 1$, for $\forall i \in N$. We call the resulting connectivity network as the farmers' information sharing network, i.e., $\mathbf{g} = \{\mathbf{g}_1, \mathbf{g}_2, \cdots, \mathbf{g}_n\}$. By definition, the network corresponds to a directed graph. Define $N_i(\mathbf{g}) = \{i \in N : g_{ij} = 1\}$ as the set of farmers (vertices) with whom i has formed a link (edge). The closure of $\mathbf{g}$ is denoted by $\overline{\mathbf{g}}$, where $\overline{g}_{ij} = max\{g_{ij}, g_{ji}\}$, which corresponds to an undirected network. Similarly, $N_i(\overline{\mathbf{g}}) = \{i \in N : \overline{g}_{ij} = 1\}$ is defined as the set of vertices directly connected to i in the undirected graph. We denote the set of all edges by $E = \{e_{ij} : \overline{g}_{ij} = 1, \forall i, j\}$, and label the vertices by $N = \{1, 2, ..., n\}$. Thus, the undirected network corresponds to the graph $G = (V, E)$. Notice that there is a one-to-one correspondence between the linking profile $\mathbf{g}$ and the induced graph G. Furthermore, we say that i and j are *connected*, if there is a path such that $\overline{g}_{iv_1} = \overline{g}_{v_1v_2} = \cdots = \overline{g}_{v_{l-1}j} = 1$, for some sequence of the vertices $v_1, v_2, \cdots, v_{l-1} \in N$. By considering the undirected closure of unilaterally formed directed network, our stylized model setup follows the economics literature, e.g., Bala and Goyal (2000). It represents an unilaterally initiated persuasion and negotiation process, which leads to the bilateral information exchange agreements.

Information structure. We use the capitalized X_i to represent the information set of farmer i, which contains all the signals that she can observe. *A priori*, her information $X_i = \varnothing$. Upon information acquisition, the information set contains her private signal

[8]For example, Moustier et al. (2010) disclose that the membership to some FPOs in Vietnam requires: (1) neighbors and kinship relationship with existing members; (2) membership fees varying from zero to 200 US dollars. In the basic model, we shall stick to the first interpretation; the latter interpretation as membership fee will be elaborated in Section 2.5.

$X_i = \{x_i\}$. In addition, the public signal is observable to all the farmers, i.e., $x_0 \in X_i$. When the information sharing network is established, her information set is updated as $X_i = \{x_i, x_0\} \cup_{j \in N_i(\bar{g})} X_j$. We abstract away from the iterative process of information updating, and assume that the ultimate information set of farmer i contains x_0, x_i, and $\{x_j\}'s$, for $\forall j$ that is *connected* with i. This stylized setup such that agents can observe their indirect neighbors' signals (global observability) captures the long-term repetitive interactions among the farmers. We assume that the communication mechanism is idealized, with no distortion, friction, or transmission loss[9]. This setting reflects the perfect information integration within an FPO and the complete separation across organizations.

Utilities. Farmer i's utility is given by

$$\Pi_i(\gamma_i, \mathbf{g_i}, q_i) = p(Q)q_i - cq_i - r\delta\{\gamma_i > 0\} - k|N_i(\mathbf{g})|, \tag{2.1}$$

depending on the information precision γ_i she chooses, the links $\mathbf{g_i}$ that she forms, and the production decision q_i. In this chapter, we use the terms *utility* and *payoff* interchangeably, while *revenue* (R_i) is referred to farmer i's utility excluding the information and connection costs.

Sequence of events. The sequence of events proceeds as follows: (1) The farmers choose the information precisions of their private signals, and decide whether to link to each other. (2) The farmers observe (the realizations of) their private signals, as well as the others' signals via the information sharing network formed. (3) Each farmer decides the production level based on their information, anticipating the rational production decisions of the other farmers. (4) The actual market price is realized and the market is cleared.

There are some caveats concerning this particular sequence of events. Firstly, the formation of FPOs and the establishment of information channels are long-term process, and should be considered before any particular production season. For example, Qiao and Yu (2013) report that each Chinese watermelon producer organization absorbs new member after a year-long probation period. Secondly, although the relative timing of network formation and information acquisition is debatable, we assume a simultaneous-move game in this layer to present an exhaustive characterization of equilibria. Thirdly, the market information reflects the long-term price trends, guiding farmers towards better production decisions. For the same reason, the information sharing process is prior to the production decision making. Our basic model does not incorporate short-term market information, which typically helps farmers to make better selling decision, e.g., when and where to sell, and not to be cheated by intermediaries. Those are interesting directions for future work.

[9]The idealized information transmission captures (at least partially) the reality, because a farmer can always pay a visit to her neighbour's field to observe what is grown and verify the market information. The readers are referred to Myatt and Wallace (2012) and the references therein for the consequences when this assumption is violated.

2.4 Analysis

Equilibrium concept. We solve the game using backward induction. In the third stage, farmer i chooses the production quantity q_i^* to maximize $\mathbb{E}[\Pi_i(\gamma_i, \mathbf{g_i}, q_i)]$. In her calculation for the expected market price, she forms an expectation of the other farmers' production levels $\mathbb{E}(q_j|X_i)$, for $\forall j \neq i$. We focus exclusively on the *linear Bayesian-Nash equilibrium*, i.e., $q_j = \overline{q}_j + \sum_{\forall x_k \in X_j} A_k^j x_k$, for some constants $\{A_k^j\}$'s. We can interpret $\overline{q}_j$ as the *base production quantity*; A_k^j as the *response factor* with respect to the signal x_k, for all signals in farmer j's information set X_j. This equilibrium concept is common in the literature, e.g., Vives (1988) and Morris and Shin (2002). When we go back to the first stage, the farmers choose the information precisions γ_i^* and the social connections $\mathbf{g_i^*}$, to maximize $\mathbb{E}[\Pi_i(\gamma_i, \mathbf{g_i}, q_i^*)]$.

A model with two farmers

We begin by analyzing a model with two farmers. The first proposition characterizes the equilibrium choices of the private information provision as well as the expected revenues for both farmers.

Proposition 1 *In the model with two farmers, no connection is made in any linear Bayesian-Nash equilibrium. For the disconnected equilibria, there are three cases depending on the information acquisition cost r, which are summarized in Table 2.1.*

Table 2.1: Equilibrium characterization for the two-farmer model.

r	Signal precisions	Expected revenues
small	$\gamma_2^* = \gamma_1^* = \frac{-2\beta+\sqrt{27\alpha^2+36\alpha\beta+13\beta^2}}{9}$	$\mathbb{E}\Pi_1 = \mathbb{E}\Pi_2 = \frac{(a-c)^2}{9b} + \frac{3}{\left(18\alpha+10\beta+4\sqrt{27\alpha^2+36\alpha\beta+13\beta^2}\right)b} - r$
medium	$\forall i = 1,2, \gamma_i^* = 0,$ $\gamma_{3-i}^* = \frac{3\alpha^2+4\alpha\beta+\beta^2}{6\alpha+4\beta}$	$\mathbb{E}\Pi_i = \frac{(a-c)^2}{9b} + \frac{\beta}{9(\alpha+\beta)^2 b}$ $\mathbb{E}\Pi_{3-i} = \frac{(a-c)^2}{9b} + \frac{9\alpha+5\beta}{36(\alpha+\beta)(2\alpha+\beta)b} - r$
large	$\gamma_2^* = \gamma_1^* = 0$	$\mathbb{E}\Pi_1 = \mathbb{E}\Pi_2 = \frac{(a-c)^2}{9b} + \frac{\beta}{9(\alpha+\beta)^2 b}$

Since the farmers can obtain private signals from external sources, it is surprising that the two farmers never want to share them. For example, consider the option for one farmer to connect and observe the other's information. In this case, the benefit for such a free-ride is dwarfed by the incentive to disconnect and obtain her own information (even if the linkage cost is much lower than the information acquisition cost). Intuitively, the information asymmetry due to isolation prevents over-reaction to common signals, reduces the quantity competition and improves the farmer's revenue.

In terms of the disconnected equilibria, the three cases depending on the information acquisition cost should be as expected: When the information acquisition is cheap, both

farmers will do so and choose an interior solution in terms of the amount of the information as a competitive outcome. If the acquisition cost is high, both farmers are satisfied with only the public signal. For the interim regime, only one farmer obtains the private signal.

Corollary 1 *When the disconnected farmers both obtain private signals, they acquire excessive private information which leads to a suboptimal aggregate payoff.*

When the private information acquisition cost is low, both farmers are better off with their private signals. In this case, it can be checked that the farmers' aggregate payoff $\mathbb{E}[\Pi_1 + \Pi_2]$ is maximized when $\gamma_1^* = \gamma_2^* = \frac{3\alpha+\beta}{9}$. The information asymmetry leads to a *competition effect*: A farmer's equilibrium private information provision increases in the other farmer's signal precision, i.e., $\frac{\partial \gamma_i^*(\gamma_{3-i})}{\partial \gamma_{3-i}} > 0$, $\forall i = 1, 2$. Consequently, they both acquire too much private information such that $\gamma_1^* = \gamma_2^* = \frac{-2\beta+\sqrt{27\alpha^2+36\alpha\beta+13\beta^2}}{9} > \frac{3\alpha+\beta}{9}$.

Corollary 2 *When the information acquisition cost r is large, the revenues increase in the public information provision. When r is small, the farmers' revenues decrease in the public information provision.*

The role of the public information provision is subtle. When the information acquisition cost is high, the farmers rely on the free public signal to make production decisions. Since we focus on the information-scarce regime, the value of the public information increases in its precision. On the other hand, when the information acquisition cost is low, the competition effect brings about over-precision in terms of private information. However, surprisingly, the public information increases farmers' incentives to obtain private signals. It can be checked that $\frac{\partial \gamma_i^*}{\partial \beta} > 0$. Intuitively, strong public information is a herding signal which reduces the farmers' reliance on their private information. Consequently, the farmers *obtain too much private information and use it too little*, which leads to lower expected payoffs.

More than two farmers

We shall proceed to discuss the general model with multiple farmers. Firstly, we restrict our discussion to the equilibria such that the information sharing network is fully connected.

Lemma 1 *Suppose that the information acquisition cost r is small. The following is true about a fully connected information sharing network.*

1. *The farmers always respond positively towards the public signal.*
2. *The expected payoffs are independent of the public information provision.*
3. *The value of information diminishes when the population size increases.*

Although the farmers always respond positively towards the public signal, this does not imply its value in coordinating farmers' production decisions. In fact, the direct informational intervention by providing public signal does not improve the farmers' revenues. Intuitively, this is because the public and the private information provisions are substitutes up to a satiation point, and the farmers' reaction via endogenous information acquisition offsets the public information provision. This is known as the *crowding-out effect* of the public information (Colombo et al., 2014).

The third observation that the information value diminishes as the population size increases, raises the question that whether a single information coalition could be a robust equilibrium outcome. The next proposition gives a negative prediction.

Proposition 2 *When the information acquisition cost r is small, at least two coalitions (including singletons) are formed in equilibrium.*

The proposition implies that a fully connected network cannot be sustained. This is a generalization to the previous result that two farmers are always isolated. The result is in sharp contrast with existing similar models, e.g., Currarini and Feri (2014) and Lee (2014), in which the connected network is sustained in equilibrium when there are more than three farmers[10]. The driving force for the difference in equilibrium predictions is that we endogenize the private information acquisition. Intuitively, the incentives to join an FPO for market information and private information acquisition are substitutes. In Currarini and Feri (2014) and Lee (2014), once an agent severs a link, she loses the observed signal to free-ride on and puts herself in a disadvantageous position due to the *competition effect*. In our model, since this agent can adjust her signal precision, she benefits from obtaining additional private information: Not only does she increase the monopoly rent on her private signal, she also puts her neighbors in a disadvantageous position by cutting off this information sharing channel.

At the other extreme, the view that there is no information sharing in the Cournot competition is pervasive in the economics literature, e.g., Gal-Or (1985) and Vives (1984). However, our next proposition presents a different perspective.

Proposition 3 *When the linkage cost k is small, an empty network with a large number of farmers cannot be sustained in equilibrium.*

In other words, the incentive of information sharing is sustained in equilibrium, which will be the driving force to form FPOs. One of the key insights from the existing models is that, a firm with exogenous private knowledge about common demand uncertainty can enjoy monopoly rent on information, so that it will not reveal such knowledge to an rival. In our

[10]Since both models require mutual agreements to form one link, they relax the equilibrium concept to *pairwise-stability*. However, the major tension remains the same, because both equilibrium concepts check deviations via unilateral one-link removals for the connected network.

model, due to multiple substitutional information channels, information sharing represents a particular sort of coalition and could influences the welfare of participating farmers favorably.

Between the two extreme cases, the next lemma characterizes the general equilibrium network configurations with multiple disconnected information sharing coalitions.

Lemma 2 *Suppose that the farmers form a network which comprises of m components, denoted as a m-way partition $(N_1, N_2, ..., N_m)$ of N, such that $N_i \cap N_j = \emptyset$, and $\cup_{i=1}^{i=m} N_i = N$. In addition, each farmer receives her own private signal x_i with precision $\gamma_i \geq 0$, as well as a public signal x_0 with precision β. In equilibrium, farmer $i \in N_i$ receives an expected payoff*

$$\begin{aligned}\Pi_i\left(N_1, N_2, ..., N_m\right) &= \frac{(a-c)^2}{(1+n)^2 b} + \underbrace{\frac{\rho_i}{\left[1+\sum_{i=1}^{m}\left(\frac{n_i\rho_i}{\alpha+\beta+\rho_i}\right)\right]^2(\alpha+\beta+\rho_i)^2 b}}_{\text{value of the private information}} \\ &+ \underbrace{\frac{\beta}{\left[1+\sum_{i=1}^{m}\left(\frac{n_i\rho_i}{\alpha+\beta+\rho_i}\right)\right]^2 b}\cdot\left[\frac{1}{\alpha+\beta+\rho_i} - \frac{\sum_{i=1}^{m}\frac{n_i}{\alpha+\beta+\rho_i}}{1+n}\right]^2}_{\text{value of the public information}} \\ &- r\delta\{\gamma_i > 0\} - k|N_i(\mathbf{g})|, \end{aligned} \tag{2.2}$$

where $n_i = |N_i|$ and $\rho_i = \sum_{j\in N_i}\gamma_j$.

Define *weak public information regime* as one in which β is sufficiently small.

Corollary 3 *The following is true about the weak public information regime.*

1. *Farmer i's payoff is quasi-concave in ρ_i.*
2. *When the i^{th} FPO obtains private information, the best-response information provision (i.e., ρ_i^*) satisfies $\frac{\partial\rho_i^*}{\partial n_i} < 0$, $\frac{\partial\rho_i^*}{\partial n_j} > 0$, and $\frac{\partial\rho_i^*}{\partial\rho_j} > 0$, for $j \neq i$.*
3. *Among symmetric equilibria where $n_i = n_j$, $\rho_i^* = \rho_j^*$, for $\forall i, j$, the aggregate payoff decreases in the number of FPOs formed and in the public information provision β.*

Consider the collective choice of signal precision ρ_i within the i^{th} coalition. The quasi-concavity suggests multiple driving forces. On one hand, the value of private information and the competition effect incentivize increasing signal precision. The competition effect is more severe when n_j is larger, for $j \neq i$. On the other hand, a farmer within the i^{th} coalition expects that the other farmers will also incorporate the same signal, which in turn *exaggerates* the production response to this signal. If n_i is larger, the exaggeration is more aggressive. We shall refer to this rationale as the *congestion effect*, i.e., the value of a private signal diminishes in the number of farmers who respond to it. The trade-off between the competition effect and the congestion effect pins down the FPO's collective choice of private

information precision. Finally, the information inefficiency due to the *competition effect* is exacerbated when the number of FPOs increases. In this case, the result documents the negative impact of the public information provision, which extends Corollary 2.

In terms of the general organization structures, we shall anticipate tree-like networks due to our sparsity regularization by costly social connections. We refer to the information sharing network as a *tree*, if the underlying graph is connected and contains no cycles. More generally, we define a *forest* as any network which contains no cycles. Intuitively, a forest consists of multiple disconnected trees.

Corollary 4 *The equilibrium information sharing network is a forest.*

A straightforward observation from Lemma 2 is that the equilibrium payoffs only depend on the information structure induced by the partitioning. A forest is the sparsest network which achieves the same coalition configurations, and farmers will not build any redundant connections.

Example 1 *(**Two-star network**) Suppose two star-shaped FPOs are formed. The set of the farmers N is partitioned into two disjoint subsets N_1, N_2 such that $N_1 \cap N_2 = \emptyset$, and $N_1 \cup N_2 = N$. The farmers in the subsets N_1 and N_2 form two connected star networks respectively. Denote $|N_1| = n_1$, $|N_2| = n_2$, and $n_1 + n_2 = n$. Furthermore, two farmers, $i \in N_1$ and $j \in N_2$, located at the centers of the stars, seek private signals x_i and x_j. Let the equilibrium signal precisions of x_i and x_j be γ_i^* and γ_j^* respectively.*

Proposition 4 *The following is true about the two-star network in the weak public information regime.*

1. *The two-star network is sustained in equilibrium, if the cost coefficients satisfy $\underline{k} < k < \overline{k}$, $\underline{r} + k < r < \overline{r}$, for some constant thresholds $\underline{k}$, $\overline{k}$, $\underline{r}$ and $\overline{r}$ (defined in the appendix).*
2. *The two FPOs will not merge when $n > 15$.*
3. *The farmers' revenues are decreasing in the public information provision β.*
4. *The equilibrium signal precisions $\gamma_i^* \propto \sqrt{\frac{n_2+1}{n_1+1}}\,(\alpha+\beta)$, and $\gamma_j^* \propto \sqrt{\frac{n_1+1}{n_2+1}}\,(\alpha+\beta)$.*
5. *The farmers' aggregate revenue decreases in the difference of group sizes $|n_1 - n_2|$.*

For the stars to be sustained in equilibrium, the costs for information sharing and direct information acquisition cannot be exorbitantly high, while the information sharing should be a more cost-effective approach than the direct information acquisition. Furthermore, the linkage cost cannot be too low, because the two stars will merge into a single connected network otherwise. In fact, the two FPOs never merge when the farmers' population is large, which can be explained by the *congestion effect*. The emergence of the star architectures also illustrates how informational leadership can arise in a setting with *ex ante* identical agents.

In terms of the farmers' revenues, the negative impact of the public information provision extends our results in Corollaries 2 and 3, and the same intuition carries through as $\frac{\partial \gamma_i^*}{\partial \beta}, \frac{\partial \gamma_j^*}{\partial \beta} > 0$. Finally, the existence of interior solutions of the private signal precisions extends the results in Proposition 1, while the fact that $\gamma_i^* \propto \sqrt{\frac{n_2+1}{n_1+1}}$ and $\gamma_j^* \propto \sqrt{\frac{n_1+1}{n_2+1}}$ where $i \in N_1$ and $j \in N_2$ further illustrates the result in Corollary 3. Finally, the farmers' aggregate revenue is maximized when the two coalitions are similar in sizes.

We have so far been analyzing the case where the public signal is weak. From now on, we shall focus on the opposite scenario, i.e., *rich public information regime*, which requires that β is sufficiently large.

Corollary 5 *In the rich public information regime, the best response information provision is chosen such that either $\rho_i \to 0$, or $\rho_i \to \infty$.*

In this regime, we focus on the value of the public information provision. Intuitively, if her group's collective private signals are much weaker than the other groups, then she will respond positively to the public signal. On the other hand, if her group's collective private signals are much stronger than the other groups, she will respond negatively to the public signal. In either case, the value of public information is high, since the public signal serves as an instrument to coordinate farmer i's production decision. We shall refer to this phenomenon as the *polarization effect* of the public information provision.

Proposition 5 ***The dominant group***. *When the cost k and r are small enough, and $r > k > 0$, the following network is formed in the rich public information regime:*

1. *Farmer $i \in N^* \subseteq N$ chooses signal precision $\gamma_i^* \to \infty$, and $\forall j \notin N^*$, $\gamma_j^* = 0$, where $\frac{n}{2} - 1 \leq |N^*| < \frac{n}{2} - \frac{1}{2}$.*
2. *$\forall j \neq i$, $j \in N^*$, either $g_{ji} = 1$, or $g_{jv_1} = g_{v_1v_2} = \cdots = g_{v_li} = 1$, for some sequence of the vertices $v_1, v_2, \cdots, v_l \in N^*$.*
3. *The rest of the farmers are isolated, i.e., $\forall j \notin N^*$, $\bar{g}_{jk} = 0$, for $\forall k \neq j$, $k \in N$.*

Following the intuitions from Corollary 5, the payoff-maximizing private signal precision should be either zero or infinity, due to the *polarization effect*. Consequently, the emerging network achieves the most *eccentric* configuration possible: The farmers who prefer strong private information will conglomerate, while those who prefer weak private information will isolate. We borrow the nomenclature *dominant group architecture* from Goyal and Joshi (2003) [11]; however, the mechanisms driving the formation of such architecture are vastly different.

[11] As the term is used in models with local observability, it requires complete connectivity among the dominant group. To be precise, the equilibrium configuration is only equivalent to the *dominant group architecture* in terms of the information structure.

Finally, we compare the farmers' aggregate revenue under the dominant group architecture, denoted as $\sum_{i=1}^{i=n} R_i^{dg}$. We further use $\sum_{i=1}^{i=n} R_i^*$ to denote the aggregate revenue socially maximized among all possible network configurations.

Corollary 6 *The following is true about the dominant group architecture.*

1. *The farmers' aggregate payoff is increasing in the public information provision β.*
2. *It is the unique class of pure-strategy equilibria among all possible coalition configurations.*
3. *For any two farmers i and j, $\lim_{n\to\infty} \frac{R_i^{dg}}{R_j^{dg}} = 1$. However, $\lim_{n\to\infty} \frac{\sum_{i=1}^{i=n} R_i^{dg}}{\sum_{i=1}^{i=n} R_i^*} = 0$.*

It is natural that the aggregate payoff increases in the public information provision, since the dominant group architecture pivots on its instrumental value in coordinating farmers' production decisions. The good news is that all farmers receive approximately the same revenue, regardless of their locations. This suggests that the dominant group architecture also achieves a fair allocation of social welfare among the farmers. This result also implies that, the accessibility of market information channels is important as it balances out the informational advantage of bigger player (the dominant FPO); consequently, the isolated farmers are not under-privileged in terms of revenues.

If there is an omniscient social planner, the best network design will be such that only one farmer obtains private information while the others rely solely on public information. From a policy perspective, *targeted information release* is required towards the implementation of social optimum. This is an interesting direction for future research. However, as the only incentive-compatible class of coalition configurations, the dominant group architecture under-performs the socially optimal network in terms of farmers' aggregate payoff. Thus, the model also contributes to our understanding of the fundamental struggle between fairness and efficiency, concerning how NGOs (and governments) should provide market information to improve welfare and eliminate poverty for farmers in developing economies.

2.5 Extensions

Heterogeneous farmers

As we interpret an individual farmer in the basic model as "the marketing board of small farmers' cooperative, or a relatively big individual farmer", her production cost reflects the size of land holdings (due to economy of scale), average education levels and production experiences (related to production efficiency), and distance to the market. Thus, we extend the basic model to incorporate the case in which the farmers are heterogeneous in terms of their production costs.

Proposition 6 *When there are two farmers with different production costs, no connection is made in any linear Bayesian-Nash equilibrium.*

The structural assumption driving this result is that the value of market information is independent of the production factors (due to linearity and additivity in the payoff function form). Consequently, the information provision affects the equilibrium outcomes in exactly the same manner as in the basic model, and all our results hold beyond two farmers. This serves as a robustness check for our major intuitions from the basic model. However, it is possible that the farmers are heterogeneous in terms of the information acquisition costs, or that the production factors are not independent of market information. This is another interesting direction for future research.

Exogenous information provision

In the basic model, we allow the farmers to obtain private signals with endogenous precisions, which depend on the farmers' search efforts. In this subsection, we assume that the farmer i's private signal precision γ_i is exogenously given. We can interpret this setting as one in which the farmers are endowed with asymmetric prior beliefs about the market condition, e.g., some villages or farmers' cooperatives may be dedicated to growing certain crop or have accumulated experience in certain niche market. The sequence of the events remains the same as in the basic model.

In general, our analysis can be applied to study all possible combinations of the private information endowment γ_i. However, to demonstrate the new insights that complement the basic model, we consider the regime with strong prior beliefs, i.e., $\frac{\gamma_i}{\beta} \to \infty$, for $\forall i \in N$, and $(\gamma_i + \gamma_j)\gamma_i > \alpha^2$, for $\forall i, j \in N$.

Proposition 7 *The following is true in the regime with strong prior beliefs:*

1. *No connection is made even if the linking cost k is zero;*
2. *This equilibrium is unique;*
3. *The farmers' aggregate payoff is maximized among all possible coalition configurations.*

When the farmers have strong prior information, they have no incentive to form any information sharing coalition. Intuitively, the benefit of FPO membership is insufficient for the farmers to relinquish their monopoly rent on the private information. The result echoes the empty network in Raith (1996), and Currarini and Feri (2014). Since the equilibrium is unique and maximizes the aggregate payoff, our results suggest a robust and global prediction. On the other hand, the results are in sharp contrast with Proposition 3 in the basic model. This sheds light on the importance of the farmers' initiatives in information acquisition to drive the formation of FPOs.

Bottom-up versus top-down

Consider the following top-down mechanism for the conglomeration of farmers' cooperatives. *A priori*, an NGO invites farmers to join FPOs at a cost k. We do not restrict the number of FPOs (m); in case where $m = n$, all the farmers are isolated. The farmers then choose the information precisions of their private signals, deciding whether to join FPO and which FPO to join. The sequence of the remaining events is the same as in the basic model: The farmers observe all the private signals within the FPO, making production decisions and the market is cleared.

In contrast, the mechanism in the basic model can be referred to as "bottom-up" approach. The FPO is formed out of spontaneous information diffusion and existing social structure. It is interesting to compare these two mechanisms. Garnevska et al. (2011) use China's example to claim that farmers' organizations are more productive and their members more active under bottom-up approach than top-down approach. Golovina et al. (2009) make similar conclusions using Russian data.

Proposition 8 *Under the top-down approach, at least two FPOs are formed in equilibrium. As the population size becomes large, complete isolation is sustained as Nash equilibrium, but not as* strong Nash equilibrium[12]*.*

Intuitively, the concept of "strong Nash equilibrium" forbids any jointly feasible and revenue-improving deviations by multiple farmers. In the network formation context, this concept is adapted to "strongly stable equilibrium", under which Lee (2014) predicts that "no information sharing" is the unique equilibrium outcome. Thus, this result stands in contrast to the existing literature.

Example 2 *Suppose that* $N = \{1, 2, 3, 4\}$*. Assume that* $\gamma_2^*, \gamma_4^* > 0$ *and* $\gamma_1^*, \gamma_3^* = 0$*. Consider a network under the bottom-up approach, wherein* $g_{12} = g_{13} = 1$*. This corresponds to the coalitions* $\{(1, 2, 3), 4\}$*, i.e., information is shared among farmer* $1, 2$ *and* 3 *but not* 4*.*

Proposition 9 *In the weak public information regime (*$\beta \to 0$*), it is possible that the configuration in Example 2 is sustained as Nash equilibrium under the top-down approach but not under the bottom-up approach.*

The key difference is due to the *information blocking effect*: Under the bottom-up mechanism, some farmers may fight against the conglomeration of FPOs. By blocking the information sharing among a larger audience, they increase the exclusiveness of their private information and possibly their revenue. We can capture this effect by the bottom-up approach, because informational hierarchy (due to their different positions in the network)

[12]This equilibrium concept follows from Myerson (1991), which states that "a strong Nash equilibrium is a Nash equilibrium such that there is no nonempty set of players who could all gain by deviating together to some other combination of strategies that is jointly feasible for them, when the other players who are not in this set are expected to stay with their equilibrium strategies."

arises naturally among *ax ante* homogeneous farmers. This effect is not captured by the top-down approach.

Proposition 10 *Suppose that the cost k and r are small enough, and $r > k > 0$. In the rich public information regime, no pure strategy Nash equilibrium exists under the top-down mechanism when $n > 2$.*

This result also stands in contrast to the bottom-up mechanism, wherein the dominant group is formed in equilibrium. To explain this result, the intuitions from the basic model still hold: The polarization effect drives some of the farmers to conglomerate while the rest are isolated. The conglomerated farmers always expect other FPO members to obtain costly private information, but never so themselves (free-riding). However, the FPO member holding private information is better off leaving the FPO due to the lack of supporting social structure. Our prediction is consistent with empirical evidence found by Golovina et al. (2009) who study top-down organized farmers' organizations in Russia. Their results are negative concerning farmers' experience and attitude towards top-down organization. Indeed, many farmers' organizations they tracked dissolve within one year.

Alternative communication protocol

Consider an alternative communication protocol referred to as *bilateral information sharing*, wherein mutual agreements are needed to build a connection such that both parties pay $k > 0$ to observe each other's signals. In contrast, we refer to the communication protocol in the basic model as *unilateral information sharing*. The sequence of events remains the same as in the basic model.

Proposition 11 *When there are two farmers under the "bilateral information sharing" protocol, no connection is made in any linear Bayesian-Nash equilibrium. When there are multiple farmers, at least two FPOs are formed in equilibrium. As the population size becomes large, complete isolation is sustained as Nash equilibrium, but not as* strong Nash equilibrium*.*

This proposition shows that the fundamental intuitions from the basic model are robust under different communication protocols. However, further results under this protocol require additional equilibrium concepts. To avoid those technical details and for clearer presentation of the major results, we stick to the communication protocol in the basic model.

Bertrand price competition

While the marketing boards of farmers' cooperatives typically make quantity decisions, in some scenarios they have gained the power of determining the prices in the market. In this

case, we study the price competition with two farmers selling differentiated products, which is an economy with complement. A farmer's payoff is given by

$$\Pi_i(\gamma_i, \mathbf{g_i}, p_i) = q_i(p_i - c) - r\delta\{\gamma_i > 0\} - k\delta\{g_{ij} = 1\}, \tag{2.3}$$

depending on her selling price p_i, information provision γ_i, and linking decision g_{ij}, for $\forall i \in \{1, 2\}$, and $\forall j \neq i$. The demand function is assumed to be

$$q_i = a + u - p_i + bp_j, \forall j \neq i, \tag{2.4}$$

where a is the market potential, $b \in (0, 2)$ is interpreted as the *cross-price elasticity*, and $u = N(0, \alpha^{-1})$ captures the market uncertainty as in the basic model. We require that $b > 0$, indicating the complementarities among the farmers, while the restriction that $b < 2$ ensures stability. We also assume the same information structure, in which the farmers can observe private signals $x_i = u + \epsilon_i$, where $\epsilon_i \sim N(0, \gamma_i^{-1})$, for $i = 1, 2$, and both farmers share a public signal $x_0 = u + \epsilon_0$, where $\epsilon_0 \sim N(0, \beta^{-1})$. The sequence of events and the equilibrium concept remain the same as in the basic model.

Proposition 12 *When β is sufficiently large, $k > 0$, no farmer will obtain private signal, and no connection is to be made. When β is sufficiently small, the farmers will be connected and choose $\gamma_1^* + \gamma_2^* = \alpha - \beta$, as long as the costs k, r are small. In both cases, the farmers' aggregate payoff is maximized.*

When the farmers' cooperatives are not saturated by the public information provision, information sharing is achieved since their incentives are aligned in a complement economy. On the other hand, strong public signal leads to social isolation. The results stand in sharp contrast to Proposition 1, wherein the two farmers' cooperatives are always disconnected in a Cournot competition. The results and intuitions remain robust when there are multiple farmers' cooperatives.

2.6 Conclusion

We propose a stylized Cournot competition model under incomplete information and study the incentives of FPOs' formation in developing economies. We focus on the functionality of FPOs as information sharing coalition. We find that there will be no information sharing between two cooperatives. In general, we should expect to observe multiple FPOs who integrate private market information within but are isolated with each other. When the public information provision is low, multiple competing FPOs are formed, and the farmers' revenues decrease in the precision of public information. On the other hand, when the public information provision is high, a fair allocation of social welfare is asymptotically achieved by the *dominant group architecture*. In this case, the farmers' revenues increase in the public information provision.

We extend the analysis in several directions to complement the basic model. We check the robustness of our results by incorporating heterogeneity in production costs, and considering alternative communication protocol. When the private information provisions are exogenous, social isolation is the unique equilibrium among farmers with strong prior beliefs. We also compare the bottom-up approach and top-down approach for FPOs' formation, and document how supporting social structure both helps the sustainability of FPOs, and deters it due to the "information blocking effect". Finally, we find that it is possible for two farmers' cooperatives to form information sharing coalition in the Bertrand price competition. This contrast to the basic model illustrates how the incentives for information sharing depend on the nature of the underlying economy.

Chapter 3

Selling Investment Goods with Present-Biased Consumers

3.1 Introduction

Background and motivation

Malaria is the most important parasitic infection in people, accounting for more than 1 million deaths a year (Greenwood et al., 2005). Combination therapy[1] with different drugs is now the preferred approach to malaria treatment. The motivation for those innovative combination treatments is driven by the emergence and spread of parasites resistant to one component of the combination; However, they are up to ten times more expensive than current mono-therapy, which are unrealistic in many settings (Whitty et al., 2004). The obstacles with respect to malaria treatment, exacerbated by the lack of efficient preventive drugs[2] and vaccines (Moorthy et al., 2004), motivate mosquito control at the centre of past efforts to eradicate malaria.

Long-lasting insecticidal nets (LLINs) are effective mosquito control strategy for sub-Saharan Africa, whose insecticidal properties last at least 4-5 years (Dupas, 2014). However, only a small percentage of individuals actually use them in most sub-Saharan countries, even though they are cost-effective in terms of investment return (Monasch et al., 2004; Hassan et al., 2008). This problem is known as technology/product adoption puzzle in developing economies, and is universal among many investment goods. For example, despite demonstrated high return of fertilizer investment (Duflo et al., 2008), fertilizer adoption rate in developing economies is low. Similar observation is made for deworming medical treatment (Miguel and Kremer, 2004), new production technology (Atkin et al., 2015), clean

[1] The most promising drugs include fixed-dose combination Dihydroartemisinin-piperaquine in southeast Asia, and combination of Artesunate and Chlorproguanildapsone designed for the African market. The discovery of Dihydroartemisinin and Artemisinin is awarded Nobel Prize in medicine in 2015.

[2] Intermittent Preventive Treatment in Infants Consortium (http://www.ipti-malaria.org).

and energy-saving cookstove (Levine et al., 2012), ceramic water filter (Guiteras et al., 2013), and so on.

In this chapter, we focus on three factors leading to the explanation of this puzzle. The first factor is due to lack of information: Farmers are uncertain about the productivity with respect to fertilizer usage (Conley and Udry, 2010), or delay adoption until observation of successful peer practice (Bandiera and Rasul, 2006). Secondly, there is usually huge heterogeneity in terms of investment return of technology in developing economies, as suggested by a field experiment in Kenya (Suri, 2011). Another empirically tested explanation for this puzzle is *present-bias effect*: Consumers have lower short-run discount factor and higher long-run discount factor. They procrastinate purchasing investment goods, but never do so later due to lack of self-control (Duflo et al., 2011). Present-bias is identified in field experiment for the adoption of insecticide treated bednets in rural India (Tarozzi and Mahajan, 2011), clean and energy-saving cookstove in Uganda (Levine et al., 2012), and ceramic water filter in Dhaka (Guiteras et al., 2013). Once this present-bias is overcome, field experiment indicates an increase in bednets adoption rate (Dupas and Robinson, 2011).

To boost the adoption of investment goods in developing economies, subsidy is the most common and accepted policy intervention. We are motivated by an environment wherein donors desire to subsidize a private and for-profit distribution channel[3]. For example, "A to Z Textile Mills" is a local manufacturer in Africa for LLINs (under royalty-free technology transfer from Sumitomo Chemical). The LLINs it produces are distributed under the financial support of the Rockefeller Foundation and ExxonMobil (Rodriguez and ole-MoiYoi, 2011). This environment appears in a wide range of healthcare markets, e.g., HIV treatment drugs (Kremer and Snyder, 2003), and vaccines (Kessing and Nuscheler, 2006). In agriculture, fertilizer subsidy in India consists of 1.52% GDP in 2008-2009 (Sharma and Thaker, 2010). In terms of the effect, subsidy is shown to be cost-effective for selling insecticide-treated nets (Cohen and Dupas, 2010). Similar affirmative evidence supporting the effectiveness of subsidy on the adoption of LLINs is found by Dupas (2014). Malawi's removal of fertilizer subsidies (due to suggested negative effects of subsidy) was followed by a famine, and the country reinstated a two-thirds subsidy on fertilizer (Dugger, 2007).

Research questions and modeling framework

As we have observed, donors are heavily involved with the distribution of investment goods through private and for-profit channel in developing economies. Throughout the chapter, we shall focus on the society's interest in boosting aggregate product adoption quantity. Specifically, we address the following research questions:

1. What are the social value of advance selling strategies in combating product adoption puzzle? Can consumer subsidy synergizes advance selling towards this aim?

[3]It is reported that donors subsidize recommended malaria drugs to commercial channels (Adeyi and Atun, 2010; Morris et al., 2015).

2. How are the effects of advance selling strategies and subsidy policies influenced by consumer attributes (lack of information, present-bias, financial conditions, etc.)?

To answer these questions, we propose a stylized monopoly pricing model with investment goods, wherein consumers suffer from *present-bias* (O'Donoghue and Rabin, 2001). Intuitively, they like myopic cash and always postpone investment for future benefit. Consequently, consumers procrastinate purchasing investment goods, but never do so later due to lack of self-control. We further assume that consumers are heterogeneous in their *sophistication level*, i.e., the degree to which they realize such present-bias. We assume a three-period horizon, wherein consumers are unaware of the product's value in the "advance-period". Their valuation for the product is heterogeneous and privately observed later in the "spot-period". The consumption benefit is generated in the future period.

We consider a seller (he) who controls the distribution channel and dictates the prices. He sets static prices for the advance- and spot-period respectively[4]. We also consider that a donor (she) desires to stimulate the purchase and use of the product. The donor represents governments, NPOs (non-profit organizations) or social entrepreneurs. The donor provides a per-unit subsidy for each purchase, while one unit of adoption generates additional payoff for her[5]. The subsidy levels are endogenous, and determined *a priori* with commitment. We investigate both scenarios wherein the subsidy is distributed either in the advance-period or in the spot-period.

This price-setting monopoly under donor subsidy is consistent with certain (but not all) operations of private for-profit distribution channel, such as Olyset® by "A to Z Textile Mills" (Rodriguez and ole-MoiYoi, 2011), and a wide range of medicines (Kremer and Snyder, 2003; Kessing and Nuscheler, 2006; Adeyi and Atun, 2010; Morris et al., 2015). In the Olyset® example, donors are companies such as Rockefeller Foundation and ExxonMobil.

Summary of results

Our model builds upon three explanations towards the product adoption puzzle in developing economies. The consumers' uncertainty for product's valuation, due to unpredictability of epidemic severity, is the driving force for strategic consumer behaviors, leading to delayed purchase. Secondly, consumers are heterogeneous in their investment return (and thus valuation), and consumers with low investment return will not adopt the prodct. Thirdly, since the benefit of LLINs is obtained in the future epidemic season, the consumers procrastinate purchase decision in the advance-period but make no purchase in the spot-period due to lack of self-control. We show that advance selling can serve as a commitment instrument in the advance-period, which is in the same spirit as the "saving account" for Philippine women (Ashraf et al., 2006). Three advance selling strategies are sustained in equilibrium, i.e., *discount*, *premium* and *no advance selling strategies*, respectively, in the order of decreasing product adoption rates.

[4]This assumption will be relaxed and dynamic pricing will be investigated in Section 3.6.

[5]This modeling framework is in the same flavor with existing literature such as Taylor and Xiao (2014).

When the consumers are less financially-constrained, and the aggregate adoption rate can be lower. Intuitively, liquidity relief measures undermine the value of advance selling instrument, and the less financially-constrained consumers are more prone to delaying good investment. Under the discount advance selling strategy, the product adoption rate increases in the severity of consumers' present-bias, while the converse is true under the premium advance selling strategy. This contrast lies in the adverse effects of present-bias on *strategic consumer behaviors* and *procrastination behaviors*, as we shall elaborate in the analysis.

When a donor desires to stimulate the product adoption by subsidizing the consumers in the spot-market, the equilibrium subsidy level increases in the fraction of the sophisticated consumers only when the present-bias is sufficiently severe, and the converse is true when the present-bias is mild. This insight guides the donor as to combine subsidy-design with increasing public awareness of their lack of self-control. Alternatively, the subsidy should be distributed to nudge the seller towards using advance selling strategy. This can be potentially more efficient, because subsidy-design synergized with advance selling strategies rewards the commitment instrument, while subsidizing direct purchase rewards procrastination and strategic consumer behaviors.

The rest of this chapter is organized as follows. Section 3.2 reviews relevant literature. Section 3.3 introduces our model setup. In Section 3.4, we carry out the analysis of seller's pricing strategies. Section 3.5 analyzes subsidy policies. Section 3.6 discusses and extends our basic model. Section 3.7 concludes. All proofs are provided in the appendix.

3.2 Literature Review

Our research is related to the literature on the product/technology adoption puzzle. The first explanation of this puzzle is due to lack of information. Conley and Udry (2010) investigate the role of social learning in the knowledge diffusion and adoption of fertilizer among Ghana pineapple farmers. Their empirical evidence implies that farmers learn from their informational neighbors who were surprisingly successful in previous periods. Bandiera and Rasul (2006) study the low adoption rate of sunflower in Mozambique. They document strategic delay in adoption to free-ride on the information gathered by others. Alternatively, Duflo et al. (2011) use present-bias to explain Kenya farmers' procrastination in fertilizer purchase in their empirical research. Similar procrastination behavior is also found in mammography treatment decisions (Fang and Wang, 2015). For healthcare products, another important dimension lies in consumers' negligence in their positive externalities. There is a general consensus that subsidizing health products with positive externalities can improve welfare (Cohen and Dupas, 2010; Kessing and Nuscheler, 2006). Miguel and Kremer (2004) and their follow-ups study a school-based deworming program in Kenya, and suggest that spillover effects alone are sufficient to justify fully subsidized deworming treatment. Atkin et al. (2015) propose a theory to explain the low adoption rate of new cutting technology among a cluster of soccer-ball producers in Sialkot, Pakistan, from the organization perspective of misalignment of incentives within firms under information asymmetry. Those effects,

exacerbated by traditional factors such as liquidity and credit constraints (Giné and Yang, 2009; Cole et al., 2013; Tarozzi et al., 2014), are leading causes of the adoption puzzle.

In developing economies, present-bias effect is particularly strong. For example, field experiment by Duflo et al. (2011) suggests that 69 percent of farmers are stochastically present-biased. We model this effect following the quasi-hyperbolic discounting framework by O'Donoghue and Rabin (2001), which is followed up by O'Donoghue and Rabin (2006). Bernheim et al. (2013) propose a standard inter-temporal allocation problem with credit constraint to investigate the casual relationship between poverty and lack of self-control. Their results imply that poverty *exacerbates* lack of self-control. This explanation both supplements and complements the theory of "whether a poor person spends relatively more of his budget on alcohol than a richer person does on designer drugs or Ipads" by Banerjee and Mullainathan (2010). Present-bias is also identified in field experiment for the adoption of clean and energy-saving cookstove in Uganda (Levine et al., 2012), and the adoption of ceramic water filter in Dhaka (Guiteras et al., 2013).

The literature towards subsidy policy is, however, mixed. Cohen and Dupas (2010) find affirmative evidence of subsidy by a randomized field experiment in Kenya, wherein malarial insecticide-treated nets are sold to pregnant women with randomized prices. Dupas (2014) uses a randomized field experiment to estimate the effects of a one-time, targeted subsidy on the long-run adoption of malarial insecticide-treated nets. Her results suggest a positive learning effect of subsidy. Mobarak et al. (2012) document the low adoption rate of non-traditional cookstoves among women in rural Bangladesh, as they do not perceive the future value in eliminating indoor air pollution and health hazard. They also find that the effect of heavy subsidy is limited (50% subsidy only leads to 12% (5%) increase in the adoption of efficiency (chimney) cookstoves, respectively). Guiteras et al. (2013) offer similar insights concerning the adoption of preventative health technologies. They offer alternative and innovative subsidy measures by combining free trial/return, and delayed payment/micro-loans.

Our work falls into the rising research agenda on socially responsible operations, initiated by Sodhi and Tang (2014). Chen et al. (2013) examine the ITC e-Choupal network and focus on its role of facilitating technology adoption/diffusion. Chen et al. (2014) further study the peer-to-peer information sharing in Avaaj Otalo, which is related to the social learning barrier towards technology adoption. This is followed up by Chen and Tang (2015), and Tang et al. (2014). An et al. (2015) study the impacts of aggregating farmers through formal or informal cooperatives, and in particular, reducing production cost by technology diffusion. Finally, the readers are referred to Lee and Lee (2007) for classic supply chain literature in developing economies. Sodhi and Tang (2014) provide a survey and future research directions in socially responsible operations.

3.3 Model

We consider a model in which a profit-seeking firm sells long-lasting insecticidal nets (LLINs) to consumers. We index time periods by $t = 0, 1, 2$, where $t = 0$ denotes the advance-period, while $t = 1$ denotes the spot-period. For example, in the case of epidemic malaria, $t = 0$ is the dry season, $t = 1$ is the rainy season when mosquitoes are breeding, and malaria season begins at $t = 2$.

Consumers. The consumer (she) population acquainted with the product is normalized to Λ_0 and Λ_1, respectively, depending on the arrival times $t = 0, 1$. The consumers who arrive in advance-period decide whether they should buy it now, postpone the purchase decision to spot-period, or simply walk away without purchase. The consumers who arrive in spot-period decide whether they should make the purchase at all. The utility of consuming the product is θV, generated at $t = 2$ (thus "investment good"). $V > 0$ measures the intrinsic value of the product, which is public and deterministic. θ measures a random consumer's uncertain valuation for the product, which is realized at $t = 2$. θ is distributed on $[0, 1]$ according to a public distribution function $F(\cdot)$, while its realization is the consumer's private knowledge. The product is durable, and there is no need for repetitive purchase. Denote a random consumer's purchase decision in period t by a_t, such that $a_t = 1$ if she makes the purchase in period t, and $a_t = 0$ otherwise.

Consider the example of LLINs. Consumers may contract malaria at $t = 2$ (epidemic season). In this example, malaria elicits a loss of $-V$, which is interpreted as the suffering of the consumer, including the cost for malaria treatment. If a consumer goes without the product, she contracts malaria in period 2 with certain probability. If she uses LLINs, she is immune to the disease at $t = 2$.

In this example, the benefit of consumption takes a multiplicative form between the private and intrinsic valuation (θ and V). This functional structure captures the interaction of a consumer's susceptibility of the disease and severity of the epidemics, wherein θ is interpreted as the probability of her contraction of malaria. θ is heterogeneous among consumers because (1) consumers' distance to the epidemic region varies, and (2) consumers' susceptibility differs. Thus, θ is realized only after a rainy season. Empirical evidence suggests that consumers' idiosyncratic attributes play a big role in their willingness of technology/product adoption. For example, a field experiment in Kenya (Suri, 2011) suggests a huge heterogeneity in terms of investment return of fertilizer applications. Similar evidence supports the heterogeneity of malaria susceptibility (Smith et al., 2005).

Quasi-hyperbolic discounting. Following the modeling framework by O'Donoghue and Rabin (2001), we assume that consumers suffer from *quasi-hyperbolic discounting*. This is a particular form of dynamic time-inconsistency, under which an individual's time discount factor between two consecutive future periods is $\delta \in (0, 1]$ but between the current period and the subsequent period $\beta\delta$, with $\beta \in (0, 1]$. The parameter δ is the traditional discount factor whereas β is called the *present-bias* factor. β measures the short-run impatience or myopia.

To be specific, in our three-period model, the present value a quasi-hyperbolic consumer

at advance-period for her aggregate future payoffs is $u_0 = v_0 + \beta \sum_{t=1,2} \delta^t v_t$, wherein v_t is the instantaneous utility flow at period t.

Furthermore, we use *sophistication* (*naivety*) to indicate the extent to which a quasi-hyperbolic consumer is (un)aware of the present-bias of her future self. In general, we parameterize the consumers by their *sophistication* level ($\hat{\beta}$), , where $\hat{\beta} \in [\beta, 1]$, following O'Donoghue and Rabin (2001). In our three-period model, the prediction of a consumer at advance-period about her spot-period utility is $\hat{u}_1 = v_1 + \hat{\beta}\delta v_2$; while a consumer at spot-period will evaluate present and future payoffs according to $u_1 = v_1 + \beta\delta v_2$. Thus, $\hat{\beta} = \beta$, $\beta < \hat{\beta} < 1$ and $\hat{\beta} = 1$ correspond to "sophistication", "partial sophistication" and "naivety", respectively. We assume that $\hat{\beta}$ is the consumer's private attribute, distributed on $[\beta, 1]$ according to a public distribution function $G(\cdot)$. The heterogeneity in terms of present-bias is known to literature and supported by empirical evidence (Ashraf et al., 2006). In the context of insecticide treated bednets, both present-bias and sophistication heterogeneity are identified by field experiment in rural India (Tarozzi and Mahajan, 2011).

Seller. We assume that a seller (he) controls the distribution channel and dictates the prices. He sets static prices P_0 and P_1 for the advance- and spot-period respectively. For a clear presentation of major results, we start with a model such that the seller has full commitment power over the prices. This assumption will be relaxed and the scenario without price commitment will be investigated in Section 3.6. In terms of modeling choice, price-setting monopoly is a tractable framework which fits in with our problem contexts. For example, Olyset® was the only LLIN to receive "full recommendation" by World Health Organization (WHO), whereas "A to Z Textile Mills" was the only producer in Africa for LLINs from 2003-2005. Indeed, this arrangement is being criticized as monopolistic (Rodriguez and ole-MoiYoi, 2011). Nevertheless, monopoly serves as a proxy and a first step towards our understanding of the economics before analyzing competitive environment, and it appears in a wide range of healthcare markets (Kremer and Snyder, 2003; Kessing and Nuscheler, 2006).

The seller's objective is to maximize his revenue by setting prices:

$$\max_{P_0,P_1} \pi = P_0\Lambda_0 \Pr(a_0 = 1) + \alpha P_1 \left[\Lambda_0 \Pr(a_0 = 0, a_1 = 1) + \Lambda_1 \Pr(a_1 = 1)\right], \tag{3.1}$$

wherein the seller's discount factor α is different from the consumers' in general. The production cost is normalized to zero. From a social responsibility perspective, we are also concerned with the aggregate *product adoption quantity*, denoted as

$$Q = \Lambda_0 \Pr\{a_0 = 1\} + \Lambda_0 \Pr\{a_0 = 0, a_1 = 1\} + \Lambda_1 \Pr\{a_1 = 1\}. \tag{3.2}$$

In these definitions, the probabilities are generated by a random draw of individual consumer, to measure the fractions of corresponding populations. At population level, we interpret them as *product adoption rates*, as they are consistent with the ratios of product adoption quantities over the corresponding population sizes.

Sequence of events. The sequence of events proceeds as follows: (1) At advance-period ($t = 0$), the seller sets prices P_0 and P_1 with commitment. A mass of population Λ_0

is acquainted with the product. The consumers observe β but not θ, and decide whether to purchase the product, or wait until the next period. (2) At spot-period ($t = 1$), a mass of population Λ_1 is acquainted with the product, and the consumers' private valuation θ is realized. The remaining consumers decide whether to purchase the product, and leave the market upon purchase. (3) In period 2, the product is consumed and the payoff is delivered.

3.4 Pricing Strategies

We begin with the analysis for consumers who arrive in period 0 and consider their purchase decisions. A consumer's instantaneous utility flow in period 2, if she has made a purchase, is $v_2 = \theta V$, while her payoff in period 1, if she makes the purchase therein, is $v_1(a_0 = 0, a_1 = 1) = -P_1$. In period 1, she evaluates present and future utilities by $u_1 = \beta\delta\theta V - P_1$, and thus, she makes the purchase in period 1 if $\theta \geq \frac{P_1}{\beta\delta V}$. This demand structure is supported by empirical evidence. For example, by a field experiment in Zambia using door-to-door marketing of a home water purification product, Ashraf et al. (2010) show that higher prices screen out those who use the product less.

In period 0, the consumer anticipates her future self in period 1 to receive $\hat{u}_1 = v_1 + \hat{\beta}\delta v_2$, which implies that she (putatively) will make the purchase in period 1 if $\theta \geq \frac{P_1}{\hat{\beta}\delta V}$. Therefore, the consumer in period 0 calculates her expected payoff according to

$$\mathrm{E}\left[u_0(a_0 = 0)\right] = \int_{\frac{P_1}{\hat{\beta}\delta V}}^{1} \max\left\{\beta\delta^2\theta V - \beta\delta P_1, 0\right\} dF(\theta). \tag{3.3}$$

Alternatively, the consumer receives $\mathrm{E}\left[u_0(a_0 = 1)\right] = \beta\delta^2 V\mathrm{E}(\theta) - P_0$ if she makes purchase in period 0. From this analysis we can clearly understand the consequences due to lack of information: Consumers strategically wait until spot-period when uncertainty in valuation is resolved. This effect is well-known in the literature as *strategic consumer behaviors*.

The analysis for a consumer who arrives in period 1 is standard: She evaluates present and future utilities by $u_1 = \beta\delta\theta V - P_1$, and thus, she makes the purchase in period 1 if $\theta \geq \frac{P_1}{\beta\delta V}$; otherwise, she simply walks away.

From now on, we make the following functional assumptions: (1) $F(\cdot)$ is a uniform distribution. (2) $G(\cdot)$ is a two-type distribution such that a fraction γ of the population Λ_0 is sophisticated ($\hat{\beta} = \beta$), while $1 - \gamma$ of which is naive ($\hat{\beta} = 1$). Under these functional assumptions, we are prepared for the following equilibrium characterizations.

Proposition 13 *Three pricing strategies are sustained in equilibrium*[6]*:*

[6]The terminology for different advance selling strategies follows classic literature, e.g., Xie and Shugan (2001). The term "premium advance selling" speaks relatively to "discount advance selling", and does not imply a high price in absolute terms.

- *Equilibrium-D, "discount advance selling": All consumers who arrive in period* 0 *make purchase in period 0 (pooling equilibrium), denoted by the superscript D;*

- *Equilibrium-P, "premium advance selling": Among those who arrive in period* 0*, sophisticated consumers buy in period 0, while naive consumers do not (separating equilibrium), denoted by the superscript P;*

- *Equilibrium-N, "no advance selling": No consumers who arrive in period* 0 *participate in the advance-selling market (pooling equilibrium), denoted by the superscript N.*

  The equilibrium prices, revenues and the product adoption quantities in pooling equilibria are summarized in Table 3.1.

strategies	discount advance selling	no advance selling
prices	$P_0^D = \frac{\beta^2\delta^2(\alpha\Lambda_1+\beta\delta\Lambda_0)\left[\alpha(4-\beta)\Lambda_1+\beta^2\delta\Lambda_0\right]V}{2(2\alpha\Lambda_1+\beta^2\delta\Lambda_0)^2}$ $P_1^D = \frac{\beta\delta(\alpha\Lambda_1+\beta\delta\Lambda_0)V}{2\alpha\Lambda_1+\beta^2\delta\Lambda_0}$	$P_0^N > \frac{(2+\beta^2)\beta\delta^2V}{2}$ $P_1^N = \frac{\beta\delta V}{2}$
revenues	$\pi^D = \frac{\beta\delta(\alpha\Lambda_1+\beta\delta\Lambda_0)^2V}{2(2\alpha\Lambda_1+\beta^2\delta\Lambda_0)}$	$\pi^N = \frac{\alpha\beta\delta(\Lambda_0+\Lambda_1)V}{4}$
adoptions	$Q^D = \Lambda_0 + \Lambda_1\left[\frac{1}{2} - \frac{\beta(2-\beta)\delta\Lambda_0}{4\alpha\Lambda_1+2\beta^2\delta\Lambda_0}\right]$	$Q^N = \frac{\Lambda_0+\Lambda_1}{2}$

Table 3.1: Characterization of the pooling equilibria.

Under premium advance selling strategy: If $\beta \leq \frac{1}{2} - \frac{\alpha[(1-\gamma)\Lambda_0+\Lambda_1]}{\gamma\delta\Lambda_0}$*, then* $P_1^P = 0$*,* $P_0^P = 0$*,* $\pi^P = 0$*. If* $\frac{1}{2} - \frac{\alpha[(1-\gamma)\Lambda_0+\Lambda_1]}{\gamma\delta\Lambda_0} < \beta < 1 - \frac{\alpha[(1-\gamma)\Lambda_0+\Lambda_1]}{\gamma\delta\Lambda_0}$*, then* $P_1^P = \beta\delta V$*,* $P_0^P = \frac{\beta\delta^2V}{2}$*,* $\pi^P = \frac{\gamma\beta\delta^2\Lambda_0V}{2}$*. If* $\beta \geq 1 - \frac{\alpha[(1-\gamma)\Lambda_0+\Lambda_1]}{\gamma\delta\Lambda_0}$*,*

$$
\begin{aligned}
P_1^P &= \frac{\beta\delta\left\{\alpha\left[(1-\gamma)\Lambda_0+\Lambda_1\right]+\gamma\beta\delta\Lambda_0\right\}V}{2\alpha\left[(1-\gamma)\Lambda_0+\Lambda_1\right]-\gamma\left(1-2\beta\right)\delta\Lambda_0},\\
P_0^P &= \frac{\beta\delta^2\left\{\begin{array}{c}\alpha\left(1+2\beta\right)\left[(1-\gamma)\Lambda_0+\Lambda_1\right]\\ -\beta\left(1-2\beta\right)\delta\gamma\Lambda_0\end{array}\right\}\left\{\begin{array}{c}\alpha\left[(1-\gamma)\Lambda_0+\Lambda_1\right]\\ +\beta\gamma\delta\Lambda_0\end{array}\right\}V}{2\left\{\alpha\left[(1-\gamma)\Lambda_0+\Lambda_1\right]-\gamma\left(1-2\beta\right)\delta\Lambda_0\right\}^2},\\
\pi^P &= \frac{\beta\delta\left\{\alpha\left[(1-\gamma)\Lambda_0+\Lambda_1\right]+\gamma\beta\delta\Lambda_0\right\}^2V}{4\alpha\left[(1-\gamma)\Lambda_0+\Lambda_1\right]-2\gamma\left(1-2\beta\right)\delta\Lambda_0},\\
Q^P &= \gamma\Lambda_0+\left[(1-\gamma)\Lambda_0+\Lambda_1\right]\left[\frac{1}{2}-\frac{\gamma\delta\Lambda_0}{4\alpha\left[(1-\gamma)\Lambda_0+\Lambda_1\right]-2\gamma\left(1-2\beta\right)\delta\Lambda_0}\right].
\end{aligned}
$$

Intuitively, a consumer makes a purchase at advance-period if she is sufficiently *sophisticated*: She understands that with high probability her future self will not make a good investment for this superior product at spot-period. Therefore, she commits herself in the

advance selling period by making the purchase early. In contrast, a sufficiently *naive* consumer *procrastinates* her purchase decision until spot-period. However, she probably will not make purchase at spot-period due to lack of self-control. The divergent purchasing behaviors are driven by heterogeneity in consumers' sophistication. This theory is consistent with an interesting observation in the field experiment by Ashraf et al. (2006): Most Philippine households report that the female controls the household finances. Because of their financial responsibilities, women are more aware of their time inconsistency and indeed are significantly more likely to open a savings account than men.

Now that we understand consumers' behaviors, we elaborate on seller's pricing strategies. We begin by observing the spot-period price:

$$P_1^P = \frac{\beta\delta V}{2}\left\{1 + \frac{\gamma\left(1-\beta\right)\delta\Lambda_0}{\alpha\left[(1-\gamma)\Lambda_0 + \Lambda_1\right] - \gamma\left(1-2\beta\right)\delta\Lambda_0}\right\} > \frac{\beta\delta V}{2} = P_1^N. \tag{3.4}$$

$$P_1^D = \frac{\beta\delta V}{2}\left\{1 + \frac{\left(2-\beta\right)\beta\delta\Lambda_0}{\beta^2\delta\Lambda_0 + 2\alpha\Lambda_1}\right\} > \frac{\beta\delta V}{2} = P_1^N. \tag{3.5}$$

The first observation is the *inter-temporal cannibalization effect*[7], driven by strategic consumer behaviors: The spot-period price is charged as an inter-temporal price discrimination instrument. Therefore, the equilibrium spot-period price is marked up from the optimality ($\frac{\beta\delta V}{2}$), causing revenue loss for the seller. The inter-temporal cannibalization effect is stronger when the present-bias is less severe, because consumers are more strategic and the mark-up under discount advance selling strategy can increase, as β increase. On the other hand, advance selling serves as a *commitment instrument*. Therefore, the existence of interior solutions of spot-period prices reflects the tension between value of commitment instrument and inter-temporal cannibalization loss. Secondly, the value of incentivizing the sophisticated consumers to commit in the advance-period increases in the severity of the present-bias, and so does the spot-period price under premium advance selling strategy. This reflects that the *value of price discrimination* increases in the severity of present-bias.

Figure 3.1 illustrates the positioning of equilibria depending on the model primitives. A low discount factor δ suggests high delay sensitivity or tight budget constraints. It is optimal for the seller to shut down advance selling channel when δ is low, as the profitability from advance sale is low. Ironically, this implies that the sale will be postponed when the consumers are most delay-sensitive. When the consumers' discount factor δ is high, premium advance selling strategy (separating equilibrium) is dominant if the present-bias is severe, as the value of price discrimination is high.

In terms of the population constitution, the premium advance selling strategy is dominant when the fraction of sophisticated consumers is high, as the value of inter-temporal price discrimination is high. Conversely, the pooling equilibria will be dominant: If the advance-period demand is high, it is optimal to serve them now than later; otherwise, it is optimal to postpone all sales in the spot-period.

[7]We follow the terminology of Parlaktürk (2012). The readers are referred to this paper and the references therein for details.

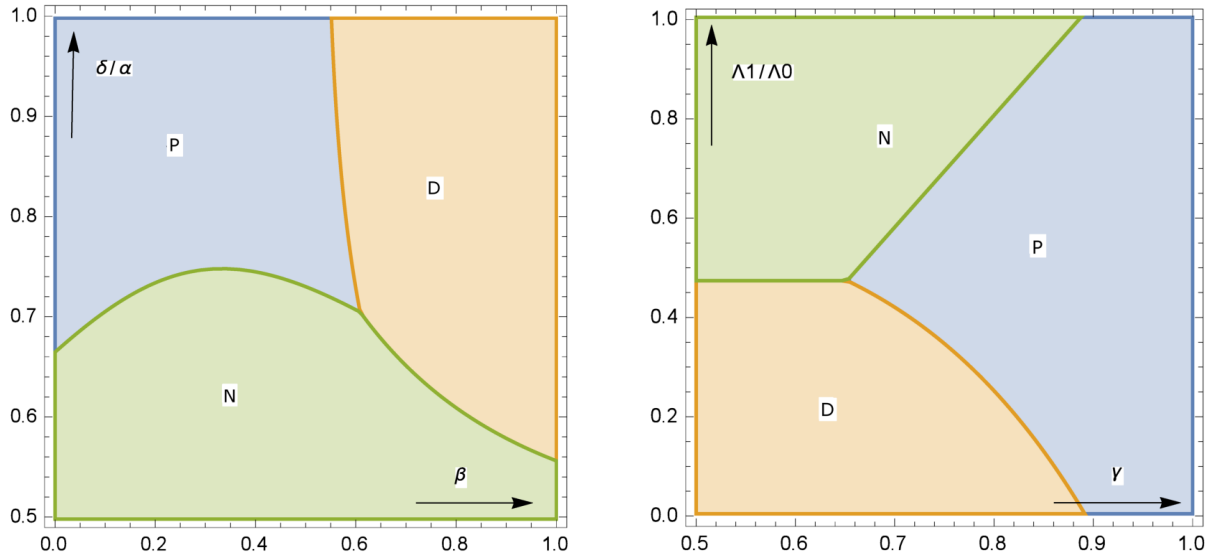

Figure 3.1: Positioning of the revenue-maximizing pricing strategies. β: degree of present-bias. $\frac{\delta}{\alpha}$: ratio of discount factors between consumers and seller. γ: fraction of the sophisticated consumers. $\frac{\Lambda_1}{\Lambda_0}$: ration of population between periods. D: discount advance selling strategy. P: premium advance selling strategy. N: no advance selling strategy. In this example, $\alpha = 1$, $\Lambda_1 = 1$, and $V = 1$. On the left-hand side, $\frac{\Lambda_1}{\Lambda_0} = 0.2$ and $\gamma = 0.8$. On the right-hand side, $\frac{\delta}{\alpha} = 0.8$ and $\beta = 0.6$.

Corollary 7 ***(Social value of advance selling)*** *Suppose that* $\beta \geq 1 - \frac{\alpha[(1-\gamma)\Lambda_0 + \Lambda_1]}{\gamma\delta\Lambda_0}$. *The following is true concerning the product adoption quantities.*

- $Q^D > Q^N$;
- *If either* $\gamma < \frac{1}{2}$ *or* $\frac{\delta}{\alpha} < \frac{2(1-\gamma)}{(2\gamma-1)\gamma}$, $Q^D > Q^P$, *for any* β, Λ_0 *and* Λ_1;
- *If* $\frac{\delta}{\alpha} < 2(1-\gamma)$, $Q^P > Q^N$, *for any* β, Λ_0 *and* Λ_1.

The aggregate product adoption quantity is always higher under the discount advance selling strategy than under no advance selling strategy. When the consumers' discount factor is low enough (time-sensitive, budget-constrained) or the fraction of sophisticated consumers is low enough, the discount advance selling strategy achieves higher adoption quantity than the premium advance selling strategy. Similarly, the premium advance selling strategy achieves higher adoption quantity than no advance selling strategy, as long as the consumers' discount factor is low enough. These results theorize the social benefit of advance selling strategies. For example, in the field experiment by Duflo et al. (2011), a field officer

would sell fertilizer to farmers immediately after harvest[8]. As farmers who participate in this program commit themselves by advance purchase, this program reduces procrastination and increases fertilizer adoption.

Corollary 8 ***(Impact of financial instruments)*** $\frac{\partial Q^D}{\partial \delta} < 0$, $\frac{\partial Q^N}{\partial \delta} = 0$. *When* $\beta \geq 1 - \frac{\alpha[(1-\gamma)\Lambda_0+\Lambda_1]}{\gamma\delta\Lambda_0}$, $\frac{\partial Q^P}{\partial \delta} < 0$.

The impact of consumers' discount factor δ on the product adoption quantities is informative concerning the effectiveness of liquidity relief measures (e.g., micro-finance instruments such as low rate loans) provided by governments or NGOs. Surprisingly, however, the product adoption quantities decrease when the consumers are less financially constrained under the advance selling strategies. Intuitively, the liquidity relief measures undermine the value of advance selling instrument. When the consumers are less financially-constrained, they are more prone to delaying good investment, and the aggregate adoption quantity may decrease. We shall elaborate on the consequences of this result on the implementation of advance selling strategies in Section 3.6.

Corollary 9 ***(Impact of present-bias)*** *Suppose that* $\beta \geq 1 - \frac{\alpha[(1-\gamma)\Lambda_0+\Lambda_1]}{\gamma\delta\Lambda_0}$. $\frac{\partial Q^P}{\partial \beta} > 0$, *and* $\frac{\partial Q^N}{\partial \beta} = 0$. *When* $\frac{\delta}{\alpha} < \frac{2(1-\beta)\Lambda_1}{\beta^2\Lambda_0}$, $\frac{\partial Q^D}{\partial \beta} < 0$.

This corollary documents the interplay between the seller's advance selling strategies and consumers' present-bias. The tension lies in the adverse effects of present-bias on strategic consumer behaviors and procrastination behaviors. Under premium advance selling strategy, the product adoption quantity increases in β (degree of present-bias). This is because the value of commitment instrument is lower when the present-bias is less severe. Consequently, the value of inter-temporal price discrimination decreases, the spot-period price decreases, and the adoption rate therein increases accordingly. Under discount advance selling strategy, the product adoption quantity decreases in β. This is because the consumers are more *strategic* (willing to wait), as the present-bias is less severe. Thus, the spot-period price increases and further mark-up from optimality is needed due to inter-temporal cannibalization. Consequently, adoption rate therein decreases.

3.5 Consumer Subsidy

Now that we understand the seller's pricing strategies in the basic model, we incorporate a donor as another stakeholder that represents the society's interest in promoting the product's adoption. In practice, donors are indispensable players in the distribution channel for investment goods in developing economies. For example, 90% A to Z's first-year produced LLINs are distributed under the financial support of the Rockefeller Foundation and ExxonMobil

[8]To ensure that short-term liquidity constraints did not prevent farmers from making a decision on the spot, farmers were offered the option of paying either in cash or in maize (valued at the market price).

(Rodriguez and ole-MoiYoi, 2011). Dupas (2014) documents the effective subsidy levels for LLINs ranging from 40% to 100%.

Donor. Suppose a donor (she) desires to stimulate the purchase and use of the product. The donor represents governments, NPOs (non-profit organizations) or social entrepreneurs. The donor provides a per-unit subsidy $s \geq 0$ to the consumers, and receives utility $W > 0$ for each purchase made. Denote the corresponding product adoption quantity as $Q(s)$, depending on the subsidy level s. The donor chooses a subsidy level s to maximize her utility $(W - s)\,Q(s)$. We assume that the subsidy level is determined *a priori*, i.e., $t = -1$. Once the donor decides, she commits to the subsidy level throughout the game. The sequence of events proceeds as before.

In general, the subsidy can be distributed both in the advance-period and in the spot-period. To isolate their separate effects, we shall elaborate either of the situations in the analysis. In addition, the definition of the donor's objective implies that she is solely concerned with the aggregate product adoption quantity. She has a flat time preference and is less financially constrained, i.e., a discount factor of one. This assumption is not crucial and can be relaxed.

Proposition 14 *(Subsidizing spot-period purchase) The following is true concerning the effect of consumer subsidy in the spot-period.*

- *(1)* $\frac{\partial Q^P(s)}{\partial s} > 0$*, if* $\beta \geq \bar{\beta}(s)$ *for some threshold* $\bar{\beta}(s)$ *depending on the subsidy level* s*;* $\frac{\partial Q^P(s)}{\partial s} = 0$*, if* $\frac{1}{2} - \frac{\alpha[(1-\gamma)\Lambda_0+\Lambda_1]}{2\gamma\delta\Lambda_0} < \beta < \bar{\beta}(s)$*. (2)* $\frac{\partial Q^D(s)}{\partial s} > 0$*. (3)* $\frac{\partial Q^N(s)}{\partial s} > 0$*.*
- *The equilibrium subsidy level increases in the degree of present-bias (i.e., decreases in* β*).*
- *Under premium advance selling strategy, the equilibrium subsidy level decreases in* γ *(fraction of the sophisticated consumers) when the present-bias is mild, and increases in* γ *when the present-bias is severe.*

When spot-period prices ($\beta \geq \bar{\beta}(s)$ case) rest in interior regions, a more generous subsidy incentivizes the seller to lower the spot-period prices, and the product adoption quantities increase under all pricing strategies; Otherwise, subsidy has no effect when the spot-period market is closed ($\frac{1}{2} - \frac{\alpha[(1-\gamma)\Lambda_0+\Lambda_1]}{2\gamma\delta\Lambda_0} < \beta < \bar{\beta}(s)$ case). The equilibrium subsidy level increases in the degree of present-bias, as it has positive effects on the marginal benefit of unit subsidy under all pricing strategies: Under no advance selling strategy, a decrease in β leads to lower perceived value of future consumption. Thus, the same amount of subsidy becomes more attractive. Under discount/premium advance selling strategy, the value of commitment increases in the severity of present-bias.

The last statement calibrates the value of inter-temporal price discrimination, due to the interplay between the consumer subsidy and present-bias. When the present-bias is mild, the commitment value is low compared to the inter-temporal cannibalization loss, and the marginal benefit of unit subsidy decreases in the fraction of the sophisticated consumers.

When the present-bias is severe, the converse is true. This result is informative concerning the value of consumer education: If the present-bias is severe, it will be helpful to increase the *public awareness* of present-bias and lack of self-control.

However, the *efficiency* of this subsidy distribution channel may be low. We measure this "efficiency" by examining consumers' share of each unit of subsidy under different pricing strategies, i.e., "who are the beneficiaries of subsidy". We can identify that

$$\begin{aligned}
\frac{s-\left[P_1^P(s)-P_1^P(0)\right]}{s} &= \frac{\alpha\left[(1-\gamma)\Lambda_0+\Lambda_1\right]}{2\alpha\left[(1-\gamma)\Lambda_0+\Lambda_1\right]-\gamma\left(1-2\beta\right)\delta\Lambda_0} < \frac{1}{2},\\
\frac{s-\left[P_1^D(s)-P_1^D(0)\right]}{s} &= \frac{\alpha\Lambda_1}{2\alpha\Lambda_1+\beta^2\delta\Lambda_0} < \frac{1}{2}.\\
\frac{s-\left[P_1^N(s)-P_1^N(0)\right]}{s} &= \frac{1}{2}.
\end{aligned}$$

In other words, advance selling strategies reduce the efficiency of subsidy distribution channel for spot-period purchase. Intuitively, subsidy distributed in the spot-market essentially *rewards* strategic consumer behaviors, and the seller shares a bigger pie of the subsidy by inter-temporal price discrimination. Alternative, we investigate the effect of subsidy for the advance-period purchase in what follows.

Proposition 15 *The following statements are true when the donor subsidizes in the advance-period.*

- *For given pricing strategy (fix an equilibrium), the product adoption quantities are independent of the subsidy in advance-period.*
- *Suppose that $\pi^D < \pi^P$ under no subsidy. If either $\gamma < \frac{1}{2}$ or $\frac{\delta}{\alpha} < \frac{2(1-\gamma)}{(2\gamma-1)\gamma}$, there exists a threshold s^{DP} such that $\pi^D(s) \gtreqless \pi^P(s)$ for $s \gtreqless s^{DP}$, respectively.*
- *Suppose that $\pi^P < \pi^N$ under no subsidy. If $\frac{\delta}{\alpha} < 2(1-\gamma)$, there exists a threshold s^{PN} such that $\pi^P(s) \gtreqless \pi^N(s)$ for $s \gtreqless s^{PN}$, respectively.*

This proposition implies that subsidy distributed at advance-period can increase the product adoption quantity by inducing the seller to change its pricing strategy; Otherwise, the subsidy has no impact on the product adoption. Instead of subsidizing purchase directly, the donor is able to *subsidize the seller's commitment instrument*. This is consistent with empirical observations. For example, Duflo et al. (2011) show that small, time-limited discounts just after harvest are most effect for fertilizer adoption.

3.6 Discussions and Extensions

Implementation Issues

Subsidizing consumers vs. subsidizing the seller. In terms of equilibrium product adoption quantities, it can be checked that these two approaches are mathematically equivalent in our stylized environment. However, they may have different effects in terms of implementation. For example, Olyset® is distributed under a voucher system (Rodriguez and ole-MoiYoi, 2011). The vouchers are sold to consumers at discount prices, and are redeemable against purchase of LLINs. In India, the controlled price at which fertilizers were sold to the farmer was paid back to the manufacturer as subsidy (Sharma and Thaker, 2010), as "it would be difficult to ensure that direct transfer of subsidy to millions of farmers is actually used by farmers for only buying fertilizer and there are no leakages in transfer of subsidy".

Subsidized return policy. To mitigate the strategic consumer behaviors due to lack of information (uncertainty in product valuation), return mechanism should be offered to the consumers. If the government mandates a return policy, it is not difficult to show that the profit-seeking seller has no incentive to follow this rule in our stylized environment. Alternatively, the donor can subsidize the return channel with a buy-back guarantee.

Proposition 16 *The following return policy implements the consumer subsidy of level s in the advance-period: The donor buys back the products at a compensation level of R to the consumers who return them at $t = 2$, where $R = \sqrt{\frac{2Vs}{\beta\delta^2}}$.*

As the donor commits to buy back the returned product at $t = 2$, consumers' perceived valuation for the product in the advance-period increases. It is interesting to note that this mechanism is potentially more efficient than direct subsidy, as $R \propto \sqrt{s}$. The seller may offer return options for other reasons, e.g., signaling high product quality (Moorthy and Srinivasan, 1995). This is an interesting direction for future research.

Other approaches. Our research is the first-step in understanding and mitigating the product/technology adoption puzzle from a socially responsible operations' perspective, and by no means exhaustive. Field experiments suggest an effective combination of free trials, return policy, delayed payment and micro-loans (Levine et al., 2012; Guiteras et al., 2013). Furthermore, advance selling is not the only way to mitigate consumers' present-bias. Alternative commitment devices include dedicated saving accounts, or a text message reminder system. The readers are referred to Bryan et al. (2010) for a summary of commitment devices.

General Distributions

In the basic model, we assume that $F(\cdot)$ is uniform and $G(\cdot)$ is a two-type distribution. In this extension, we consider general distributions over these consumers' attributes. We extend our analysis starting from consumers' behaviors.

Lemma 3 *When $G(\cdot)$ is a general distribution over the support $[\beta, 1]$, $\exists \beta^*$ and $\beta^{**} \in [\beta, 1]$ ($\beta^{**} \leq \beta^*$) such that consumers who arrive at advance-period strictly prefer purchase immediately if $\hat{\beta} < \beta^{**}$, strictly prefer postponing purchase if $\hat{\beta} > \beta^*$, and indifferent if $\beta^{**} \leq \hat{\beta} \leq \beta^*$.*

In other words, our intuition that "sophisticated consumers buy early and naive consumers procrastinate" still holds under general distributions. Since the seller can charge slightly lower prices $P_0 - \epsilon$ and $P_1 - \epsilon$ (taking $\epsilon \to 0$), we assume that the indifferent consumers always make the purchase by default without loss of generality. Given this understanding of consumers' behaviors, the seller maximizes profit:

$$\max_{P_0, P_1} \pi = G(\beta^*)\Lambda_0 P_0 + \alpha \left[\bar{G}(\beta^*)\Lambda_0 + \Lambda_1\right] P_1 \bar{F}\left(\frac{P_1}{\beta\delta V}\right). \tag{3.6}$$

We follow the nomenclature in the basic model concerning equilibrium-D or -N, when all or none advance-period demand is satisfied by advance selling. Equilibrium-P denotes the case when consumers at advance-period buy immediately only if she has a sophistication level $\hat{\beta} \leq \beta^*$.

Proposition 17 *Suppose that the distribution $F(x)$ (1) is continuously differentiable with density function $f(x)$; (2) has full support on $[0, 1]$; (3) has non-decreasing failure rate, i.e., the hazard rate function $H(x) = \frac{f(x)}{\bar{F}(x)}$ is non-decreasing in x; (4) and the function $x^2 f(x)$ is non-decreasing. The following statements hold for interior solutions of P_1^D, P_1^P, P_1^N, and corresponding adoption quantities.*

- *$P_1^N < P_1^D$, and $P_1^N < P_1^P$;*
- *$\lim_{\beta^* \to 0} P_1^P = P_1^N$, and $\lim_{\beta^* \to 1} P_1^P = P_1^D$;*
- *Compared with Q^N, both Q^D and Q^P have lower adoption rates among Λ_1;*
- *Compared with Q^D, Q^P has a lower adoption rate among Λ_0.*

Under advance selling strategies, the spot-period prices are marked up due to intertemporal cannibalization. Consequently, the seller increases the adoption rate among advance-period population, at the cost of lower adoption rate among the spot-period population. The two pooling pricing strategies can be viewed as the extreme cases of the premium advance selling strategy. Thus, by providing a discrimination instrument, this separating pricing strategy is a flexible framework for demand rationing across time dimension.

Heterogeneous Present-Bias

In the basic model, we assume that consumers have homogeneous shot-run time-sensitivity (β), i.e., the degree of present-bias. As the literature has documented heterogeneity in

this time-sensitivity along this dimension in developing economies (Ashraf et al., 2006), we incorporate this feature as a robustness check for our basic model.

We assume that a fraction ρ of the consumer populations (both Λ_0 and Λ_1) demonstrate high short-run discount factor β_H, i.e., they are patient or less budget-constrained (H-type), while the rest $1-\rho$ exhibit low short-run discount factor β_L (L-type). Following the quasi-hyperbolic discounting approach, we assume that all consumers are *partially naive*, and anticipate a future short-run discount factor $\hat{\beta} \geq \beta_H$. The next proposition provides an equilibrium characterization. We follow the terminology in the basic model concerning equilibrium-D or -N, when all or none advance-period demand is satisfied by advance selling. Equilibrium-P denotes the case when H-type consumers buy in advance-period, while L-type consumers do not.

Proposition 18 *Suppose that* $\frac{\beta_L}{\beta_H} < 1 - \frac{\rho}{2}$, $\hat{\beta} > \frac{1}{1+\sqrt{1+1/\beta_H}}$ *and the ratio* $\frac{\delta}{\alpha}$ *is small enough. The seller's revenues in different equilibria are given as follows:*

$$\begin{aligned}
\pi^D &= \frac{\beta_L\beta_H\hat{\beta}^2\delta(\beta_L\delta\Lambda_0+\alpha\Lambda_1)^2V}{4\alpha\hat{\beta}^2\left[(1-\rho)\beta_H+\rho\beta_L\right]\Lambda_1-2\beta_L^2\beta_H(1-2\hat{\beta})\delta\Lambda_0},\\
\pi^N &= \frac{\alpha\beta_L\beta_H\delta\left(\Lambda_0+\Lambda_1\right)V}{4\left[(1-\rho)\beta_H+\rho\beta_L\right]},\\
\pi^P &= \frac{\beta_L\beta_H\hat{\beta}^2\delta\left\{\beta_H\delta\Lambda_0\rho+\alpha\left[\Lambda_0(1-\rho)+\Lambda_1\right]\right\}^2V}{4\alpha\hat{\beta}^2\left[\beta_H\left(\Lambda_0+\Lambda_1\right)(1-\rho)+\beta_L\Lambda_1\rho\right]-2\beta_H^2(1-2\hat{\beta})\beta_L\delta\Lambda_0\rho}.
\end{aligned}$$

Additional assumptions are natural: It requires that differentiation in present-bias is significant (the ratio $\frac{\beta_L}{\beta_H}$ is small), the naivete level is lower-bounded, and the seller is less liquidity-constrained. Figure 3.2 illustrates the positioning of dominating equilibria depending on the model primitives.

The positioning of no advance selling strategy and discount advance selling strategy depends on the market size ratio Λ_1/Λ_0, as in the basic model. Under premium advance selling strategy, the option of advance purchase is more attractive for H-type consumers, while the consumers who suffer most from present-bias (L-type) do not benefit from advance-selling. Thus, the value of inter-temporal price discrimination increases in the fraction of H-type consumers. When the consumers' sophistication level decreases (higher $\hat{\beta}$), the dominating region of no advance selling strategy expands, as the value of commitment decreases. This intuition is consistent with that in the basic model.

Corollary 10 *If* $\frac{\Lambda_1}{\Lambda_0} < \frac{1}{2}$, *and* $\frac{\beta_H}{\beta_L} > \frac{2\rho^2}{(1-2\rho)^2}$, $\lim_{\frac{\delta}{\alpha}\to 0}\frac{\partial Q^P}{\partial\rho} > 0$; $\hat{\beta} > \frac{1}{2} \Leftrightarrow \frac{\partial Q^D}{\partial\rho} > 0$; $\frac{\partial Q^N}{\partial\rho} = 0$.

Under advance selling strategies, product adoption quantities increase in the fraction of $H-$type consumers, as they make use of the commitment instrument and buy early. These results complement our basic model on the impact of present-bias, and the benefit of increasing consumer awareness of their lack of self-control by means of financial responsibility education.

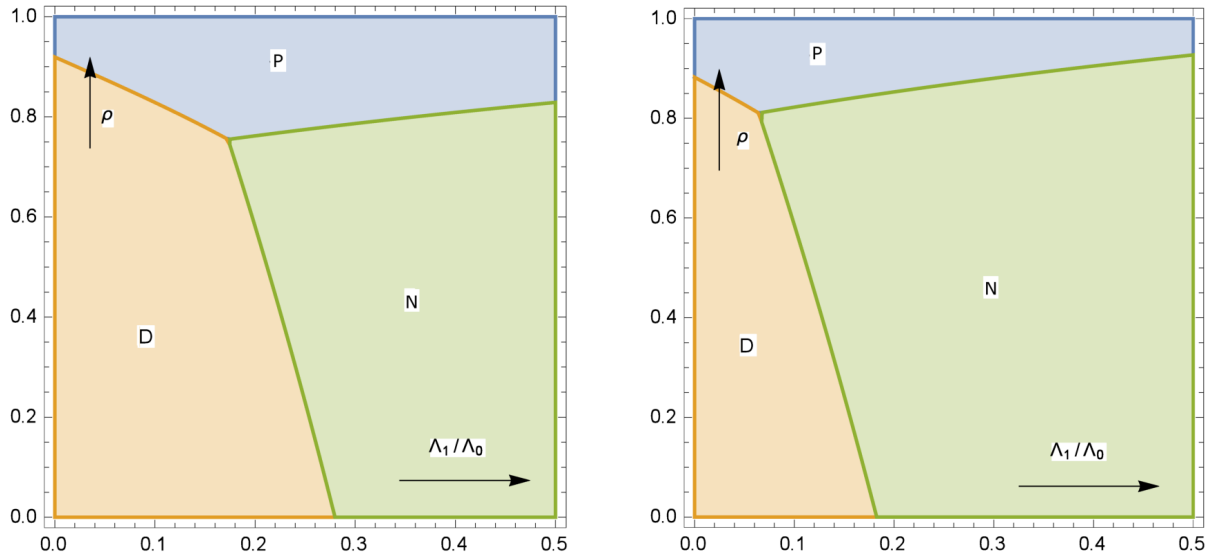

Figure 3.2: Positioning of the revenue-maximizing pricing strategies. Left-hand side: consumers are more sophisticated. Right-hand side: consumers are more naive. ρ: fraction of consumers with high short-run discount factor. $\frac{\Lambda_1}{\Lambda_0}$: ration of population between periods. D: discount advance selling strategy. P: premium advance selling strategy. N: no advance selling strategy. In this example, $\alpha = 1$, $\Lambda_0 = 1$, $V = 1$, $\beta_H = 0.7$, $\beta_L = 0.4$, $\delta = 0.7$. On the left-hand side, $\hat{\beta} = 0.7$. On the right-hand side, $\hat{\beta} = 0.9$.

Dynamic Pricing

If the seller lacks the commitment power over prices, the spot-period price is determined to maximize the spot-period revenue myopically, while the sequence of other events is identical to the basic model. We refer to this alternative environment as dynamic pricing, and characterize the equilibria in Proposition 19. Following the nomenclature in the basic model, we characterize equilibrium$-D, -P$, and $-N$. We denote the equilibrium solutions under dynamic pricing with a tilde mark.

Proposition 19 *Under dynamic pricing, the spot-period prices* $\tilde{P}_1^D = \tilde{P}_1^P = \tilde{P}_1^N = \frac{\beta\delta V}{2}$. *The advance-period prices, revenues and the product adoption quantities are summarized in Table 3.2. Furthermore,* $\tilde{Q}^D > \tilde{Q}^P > \tilde{Q}^N; \tilde{Q}^D > Q^D$, $\tilde{Q}^P > Q^P$, *and* $\tilde{Q}^N = Q^N$.

Suppose that a donor (she) desires to stimulate product adoption, as in the basic model. It can be checked that the effects of subsidy in the advance-period are similar with that under price commitment.

Prices	Revenues	Adoption Quantities
$\tilde{P}_0^D = \frac{1}{8}\beta^2(4-\beta)\delta^2 V$	$\tilde{\pi}^D = \frac{\beta\delta\{(1+2\beta)\delta\gamma\Lambda_0+2\alpha[(1-\gamma)\Lambda_0+\Lambda_1]\}V}{8}$	$\tilde{Q}^D = \Lambda_0 + \frac{\Lambda_1}{2}$
$\tilde{P}_0^N > \frac{1}{8}\beta(1+2\beta)\delta^2 V$	$\tilde{\pi}^N = \frac{\alpha\beta\delta(\Lambda_0+\Lambda_1)V}{4}$	$\tilde{Q}^N = \frac{\Lambda_0+\Lambda_1}{2}$
$\tilde{P}_0^P = \frac{1}{8}\beta(1+2\beta)\delta^2 V$	$\tilde{\pi}^P = \frac{\beta\delta[(4-\beta)\beta\delta\Lambda_0+2\alpha\Lambda_1]V}{8}$	$\tilde{Q}^P = \gamma\Lambda_0 + \frac{(1-\gamma)\Lambda_0+\Lambda_1}{2}$

Table 3.2: Equilibrium characterization under dynamic pricing.

Proposition 20 *The following statements are true concerning the marginal effects of spot-period subsidy: (1)* $\frac{\partial\tilde{Q}^P(s)}{\partial s} > \frac{\partial Q^P(s)}{\partial s}$ *if and only if* $\beta > \frac{1}{2}$*; (2)* $\frac{\partial\tilde{Q}^D(s)}{\partial s} > \frac{\partial Q^D(s)}{\partial s}$*; (3)* $\frac{\partial\tilde{Q}^N(s)}{\partial s} = \frac{\partial Q^N(s)}{\partial s}$*. The same relations hold for the equilibrium spot-period subsidy levels.*

Intuitively, due to lack of price commitment, the inter-temporal price discrimination power is weaker under dynamic pricing. Consequently, the mark up in the spot-period optimal price is removed, and the adoption quantities are higher under dynamic pricing. As the seller benefits less from subsidy, the efficiency of subsidy distribution channel is higher. Thus, the donor has a higher willingness to donate, and the equilibrium spot-period subsidy levels are weakly higher.

3.7 Conclusion

We propose a stylized monopoly pricing model with investment goods, wherein consumers suffer from *present-bias*. Consumers are heterogeneous in their *sophistication level*, i.e., the degree to which they realize such present-bias. Consumers are uncertain about product value at advance-period, which is heterogeneous and privately observed later at spot-period. Since the benefit of investment goods is generated in the future, the consumers procrastinate purchase decisions in the advance-period but fail to commit to their purchasing plans in the spot-period.

Our results are driven by the dual effects of present-bias, as it encourages procrastination behaviors while discourages strategic consumer behaviors. We show that advance selling can be beneficial both to the seller as an inter-temporal discrimination instrument, and to the consumers as a commitment device. When the consumers are less financially-constrained, the aggregate adoption rate can be lower. Under *discount advance selling strategy*, the aggregate product adoption rate increases in the severity of consumers' present-bias, while the converse is true under *premium advance selling strategy*. When a donor desires to stimulate the product adoption by subsidizing the consumers in the spot-market, the equilibrium subsidy level increases in the fraction of the sophisticated consumers only when the present-bias is sufficiently severe, and the converse is true when the present-bias is mild.

Finally, we discuss implementation issues of advance selling and different subsidizing strategies. We also show that our structural results concerning the value of advance selling

strategy are robust under general distributions over consumers' attributes, heterogeneity in terms of consumers' time-sensitivity, and dynamic pricing. These extensions also generate new insights that complement our basic model.

Chapter 4

Revenue-maximizing Pricing and Scheduling Strategies in Service Systems with Horizontal Substitutions

4.1 Introduction

The literature on scheduling and priority pricing in service systems focuses almost exclusively on the vertical dimensions of customers' heterogeneity, e.g., *willingness to pay* and *willingness to wait*. In this chapter, we extend the literature to investigate customers' horizontal heterogeneity, and we propose a model that incorporates their *taste preferences*. In particular, we allow ties in preference rankings among multiple service options. Thus, the issue of *taste indifference* arises when they are indifferent among the choices in the service menu, and a customer is *flexible*, if she demonstrates such indifference in her taste preference. The service provider, in response, can leverage on those customers and offer services with *horizontal substitutions*. This chapter investigates the value and challenges in providing services with horizontal substitutions.

The horizontal substitution problem is primarily motivated by fresh-product delivery service operations. Horizontal substitutions are headaches for grocery delivery service providers, as it is not uncommon for a specific grocery item to run out of inventory. For example, Walmart Grocery will substitute unavailable items with similar counterparts unless customers opt out substitutions. Similar service providers include Amazon Fresh, Google Express, and Instacart. The service mechanism of Instacart is closest to this chapter: It commits to an expected waiting time (2 hours minimum) with corresponding delivery fees. Then, customers make choices and pay online. The service provider schedules delivery based on certain service discipline and sends out shoppers. Once the customers finalize the deal online, they cannot renege. In particular, Instacart enables their customers to be contacted for confirmation concerning substitutions. Obviously, this takes more time for each orders and experienced shoppers are hired to fulfil customers' varying preferences. Alternatively, customers can also

specify indifference between features such as brands, origins, and etc.

Similar service mechanism is offered in food-delivery industry. For example, SpoonRocket offers food delivery service with the aid of smart phones and GPS locations. Similar companies include Seamless, GrubHub, Postmates and Sprig. In SpoonRocket's service mechanism, a contract is specified on their websites consisting of the items, e.g., tuna or kale, the respective prices, and the expected waiting times. The dedicated customers in this context are the meat-lovers who strictly prefer tuna over kale, and the vegetarian customers who strictly prefer kale over tuna. The flexible customers are those who are indifferent between ordering tuna or kale.

The presence of flexible customers is not isolated incidence in the fresh-product delivery industry. Horizontal substitution strategy for flexible customers is pervasive across a wide variety of industries. For example, in the call center service operations, polylingual customers can be served via different languages. The consideration of customer flexibility is increasingly crucial in the telecommunication industry in the United States, due to its diversifying demographical landscape. For e-commerce marketplace such as Amazon, a flexible customer could be indifferent between purchasing a white T-shirt and a black T-shirt, given that their qualities are the same. Another application is the service operation of electric vehicle charging stations. A flexible customer in this example could be a vehicle equipped with a battery system that is compatible with multiple service platforms.

The information structures in such service operations vary across different applications. For example, the compatibility of the battery systems is publicly observable while the language ability of customers is not. The information asymmetry with respect to customer flexibility is important, because the system performance varies under different information structures. When such information asymmetry exists, the service provider could adopt discriminatory service mechanisms to elicit customers' private information. By taking advantage of customer heterogeneity, discriminatory mechanisms achieve better performance in terms of revenue maximization, and they are widely used in practice. For example, Amazon provides a menu of delivery services with different lead times and charges; similar applications include the fast-pass in the traffic system. In the call center example, the delays for service vary across languages, which could be due to the staffing choice or driven by unpredictable demand.

This chapter provides a holistic analysis for the design of revenue maximizing policies for service systems when the customers demonstrate weak taste preference, and explores the impact of different information structures and the discriminatory mechanisms on the system performance. We consider a service provider who operates two different queues, e.g., one queue is for tuna delivery and the other is for kale. The customers' delay sensitivity is discretized into two classes, which correspond to impatient and patient customers respectively. There are dedicated as well as flexible customers in both queues, each served by a single server.

A more subtle issue arises: Should the service provider treat the flexible customers differently *ex post*? If the answer is no, then the flexible customers will self-select which queue to join. In the food delivery example, once an indifferent customer chooses tuna over kale,

she enjoys the same service priority as the strict meat-lovers. If the answer is yes, the service provider could probabilistically route the flexible customer to a particular queue, in which she is assigned different priority than the dedicated customers. In the e-commerce application, "opaque selling" corresponds to this idea. A customer choosing the option of opaque selling is essentially revealing her horizontal (taste) indifference, and the seller could incentivize this option by price discount or higher delivery priority. To address this issue, we compare the basic model where the service contracts are differentiating with respect to the customers' flexibility, with an alternative model without such differentiation.

Now we shall briefly preview the main results in this chapter. First, following the *revelation principle*, we propose a direct revelation mechanism that consists of six separating contracts, depending both on customers' delay sensitivity and on their flexibility. In the presence of information asymmetry, the revenue maximizing mechanism shall accommodate customers' incentive compatibility. This is in line with Mendelson and Whang (1990), Afeche (2013), and Maglaras et al. (2013). We illustrate the structures and features of the jointly optimal pricing, scheduling and routing policies. When the traffic to both queues is not balanced, the flexible customers should be assigned shorter expected delays than the dedicated ones. On the other hand, with balanced traffic inputs, the flexible customers should be assigned longer expected delays. It is possible that the dedicated customers suffer from longer delays with the increase of the flexible customers. This is different from the results shown by Akgun et al. (2012), which dictate that the dedicated customers always benefit from the increasing fraction of the flexible ones. Our analysis therefore documents the crucial role of information asymmetry on the interplay between the flexible customers and the dedicated ones.

Second, we propose a *server-specific* mechanism, in which once a flexible customer joins any of the two queues, she is treated in the same way as the dedicated customers in terms of priority. In other words, since routing is endogenized in the customers' decision process, the customers are heterogeneous only in terms of delay sensitivity *ex post*. This restriction makes the service provider lose some of the discrimination power in this *server-specific* mechanism, compared with the fully separating basic model. However, it allows us to solve the scheduling problem for given arrival rates in closed forms. The solution method for the scheduling problem follows the feasible region approach, e.g., Afeche (2004), and Yahalom et al. (2006). When one queue accommodates a large population of impatient customers, it is possible to strategically idle the server in the other queue, even if there might be awaiting patient customers in the queue. This phenomenon is new to the literature as the existing papers focus exclusively on a single-server system wherein strategic delays take place *within* the same queue. We show by numerical examples that the revenue gap between the basic model and the server-specific model is small, while the latter mechanism is easier to implement.

We further explore the value of information, and discuss the impacts of different information structures on service systems. We emphasize on the information asymmetry in the basic model, since the *server-specific* model is a special case in terms of information structures. We find that the discriminatory mechanism (as in the basic model) could increase the total revenue, and the service provider needs to pay two levels of information rent, both in

terms of the delay sensitivity and the flexibility. As long as the information about the delay sensitivity is public, the first-best could be restored (defined in Section 4.3). However, if the flexibility is observable while the delay sensitivity is not, the service provider still needs to pay information rent to the patient customers. This provides some guidelines for the service provider regarding the collective design of service rules and information acquisition if possible.

The chapter is organized as follows. In Section 4.2 we present a literature review. We give a mathematical representation of the basic model described above in Section 4.3, and provide a comprehensive analysis to the model. In Section 4.4 we consider the *server-specific* model. We summarize the main idea of this chapter in Section 4.5. The proofs for the main results, as well as additional proofs and results are provided in the appendix.

4.2 Literature Review

Our models are related to the growing literature on flexible service systems. Since this line of research has long root in the operations research literature, we refer the reader to Gans et al. (2003) for an overview of earlier papers, where the flexibility structure in our model is denoted as *W-design*. For partially pooled systems in general, Tekin et al. (2009) compare the performance across different designs, and provide insights for cross-training decisions. In call center applications, the *M-design* could arise due to clustered organization structure. For example, the *out-portfolio flow* of customers in Jouini et al. (2008) could be served by all agents, which are similar to the flexible customers in our model, while the *in-portfolio flow* are analogous to the dedicated customers in our model. From this perspective, our model relaxes their assumption that the flexible customers always suffer lower priority than the dedicated customers.

More recently, Bassamboo et al. (2012) study the effectiveness of sparse flexibility in the queueing setting, by using fluid and diffusion approximations. While almost all related papers along this line of literature focus on the *resource flexibility*, the literature devoted to customer flexibility is scarce. He and Down (2009) conclude that in many cases it is good to accommodate customer flexibility with little cost. They also raise the question of whether enough incentives can be built in to encourage enough customers to be flexible and thus allow all customers to reap the benefit. Akgun et al. (2011, 2012) adopt the stochastic comparison approach to show that the dedicated customers always benefit from the increasing fraction of the flexible customers. We tackle the problem under a mechanism design framework, which enables us to study the problem with information asymmetry and obtain richer results. In particular, the dedicated customers could be hurt when there are flexible customers who enjoy higher priorities in the system.

Our work also extends the literature on service systems with rational agents. Naor (1969) is among the earlier researchers investigating the pricing problem in service systems. In this stream of literature, agents respond to economic incentives. Classical results in both observable and unobservable queues are reviewed in Hassin and Haviv (2003). Under

their taxonomy, our model falls into the category of incentive compatible priority pricing models with unobservable queue lengths. The idea of incentive compatible priority pricing with information asymmetry starts from Mendelson and Whang (1990), which analyzes exponential systems where customers differ in terms of their willingness to pay for one unit of service and their delay sensitivity. Their results show that the famous "$c\mu$ rule" is a revenue maximizing and incentive compatible policy, where c is the delay cost rate and μ is the mean service rate. Ha (1998) studies optimal pricing in service systems with $GI/GI/1$ queue, where customers could choose service rates. The paper adopts the incentive compatible pricing schemes and points out that the service provider should reimburse customers for their actual delay cost. Ha (2001) proposes a set of two-part linear pricing schemes to coordinate the admission and service rates in the system.

More recently, Afeche (2004, 2013) extend the literature beyond the $c\mu$ rule, and introduce a family of optimal policies with strategically inserted idleness. The intuition behind the policies is that the service provider may have a strong incentive to stay idle, anticipating the future arrival of customers with higher priority. Cui et al. (2009) propose probabilistic admission control policies, under which a customer might be probabilistically rejected depending on her delay sensitivity while another customer with the same valuation might be admitted. They adopt discrete customer valuation and patience, and the resulting two-dimensional screening problem can be solved by second-order conic program. Maglaras et al. (2013) build upon the formulation of Afeche (2013); they endogenize the separation of service classes, and extend the setting to the multi-server or multi-type systems. They find that strategic delays are not the first-order effect in two-type systems where the resource constraint is binding. We adopt a similar framework with this stream of literature, but incorporate the novel feature of flexible customers. This new feature gives rise to additional complications in the optimal control policies, but also offers new insights to the mechanism design problem. Zhao et al. (2012) propose a model of make-to-order production with differentiated lead time and price quotation. However, their model and the literature therein referred to, do not incorporate customers' taste heterogeneity in the horizontal dimension. Afeche and Pavlin (2015) extend Afeche (2004, 2013) by screening both in terms of customers' patience and willingness to pay. In particular, they assume that patient customers are willing to pay more for long lead times than impatient ones, and vice versa for speedier service. While they explore along the vertical dimension of customers' preference, we focus on their horizontal differentiation.

Finally, our research is related to the literature on production and service systems with horizontal differentiation. So (2000) and Cachon and Harker (2002), are among the earliest researchers to consider differentiated services, but they focus on the oligopoly competitions in price and service time. In So (2000), the demands are generated by an exogenous attraction model. Cachon and Harker (2002) consider a queueing game, assuming exogenous logit demand functions. Allon and Federgruen (2007) investigate service competition in both service levels and prices with a general class of asymmetric demand functions.

In the production-inventory literature, Mendelson and Parlaktürk (2008) and Xia and Rajagopalan (2009) explore the leadtime-variety tradeoff in competitive settings. In Alptekinoglu

and Corbett (2010), a locational choice model is proposed to integrate product line design problem that involves variety, leadtime (or inventory), and pricing decisions. Cattani et al. (2010) address the issue of how should limited capacity be managed for make-to-stock (standard) and make-to-order (custom) production. The trade-off in this scenario echoes the literature on service systems with flexible server. In a slightly different context, Kohlberg (1983) introduces congestion (waiting costs) to the Hotelling model, i.e., location model with horizontal taste heterogeneity. Ahlin and Ahlin (2013) extend the Hotelling model with differentiated product, and show that congestion effects mitigate competition, eliminating aggressive pricing equilibria. Along this line of research, Yang et al. (2013) generate the asymmetric demand model by a *spokes model*, i.e., a location model with taste heterogeneity as an extension to the Hotelling model. Our work further differs from this literature by incorporating private taste preference in a centralized system, which also admit a non-zero mass of indifferent customers.

4.3 The Basic Model

Model formulation

We consider a capacity-constrained queueing model, where a service provider (he) employs two servers to serve heterogeneous customers (she). We adopt $M/M/1$ model for each queue, where the service time in each queue is independently exponentially distributed with a common service rate μ.

Customers. Customers have the same valuation for service, but differ in their taste and delay sensitivity. The delay sensitivity is characterized by unit delay cost C_i, $i \in \{H, L\}$, where we assume that time is more valuable for the H-type customers than the L-type ones, i.e., $C_H > C_L$. Intuitively, C_H refers to impatient customers and C_L refers to relatively patient customers. The taste preference is characterized by the customers' willingness to pay for the two servers, which are interpreted as two horizontally differentiated choices. We denote a type-k customer's willingness to pay for server m by $v(m|k)$, for $\forall k \in \{1, 2, f\}$ and $\forall m \in \{1, 2\}$. We assume that for the dedicated customers ($k \in \{1, 2\}$), $v(m|k) = V$, for $\forall m = k$, while $v(m|k) = 0$, for $\forall m \neq k$. The flexible customers have the same willingness to pay for both servers, i.e., $v(1|f) = v(2|f) = V$. We assume that V is large enough to avoid negative optimal prices.

It could be easily checked that, under this preference structure, a dedicated customer will only choose the server that matches her taste, while a flexible customer might choose either server. This capture the idea in the food delivery example that, a meat-lover will strictly prefer tuna over kale while a vegetarian will strictly prefer kale. The flexible customers might choose either meal, depending on the prices and expected delay. In what follows, we shall use *taste* and *flexibility* interchangeably. Thus, it suffices to assume that the customers arrive according to Poisson processes with aggregated arrival rates of λ_i^k, where $i \in \{H, L\}$ denotes the customer types in terms of the delay sensitivity, and $k \in \{1, 2, f\}$ denotes the

dedicated customers' arrival to the first, second queues and the arrival of flexible customers respectively. The arrival rate λ_i^k is fixed/exogneous for each customer class, $\forall i \in \{H, L\}$, and $\forall k \in \{1, 2, f\}$.

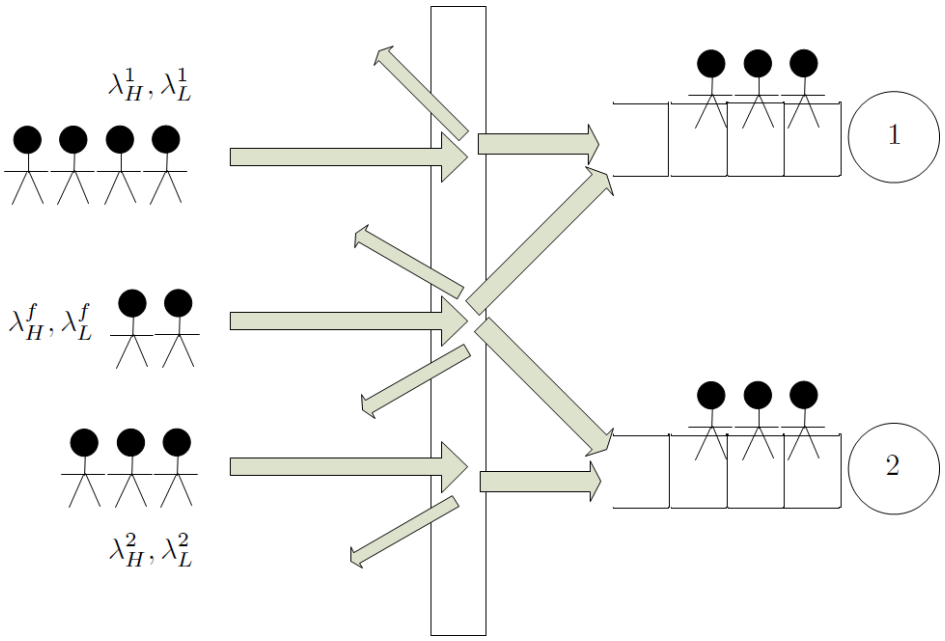

Figure 4.1: Illustration of the service system.

Information structure. In compliance with the literature on incentive compatible priority pricing, we assume that the arrival processes, service rate, the customers' willingness to pay, the cost distributions and the service procedure are common knowledge. However, a customer's delay sensitivity and the flexibility/taste are privately known to this customer and unobservable to the service provider. Furthermore, we assume that queue lengths are not observable to customers, which is motivated by the practice in the food delivery industry. This assumption is common in the literature (Hassin and Haviv, 2003).

Service mechanism. The service provider tries to design an appropriate mechanism to maximize his long-run average payoff. Since the customers' attributes of delay sensitivity and flexibility/taste are both private information, the service provider faces an adverse selection problem and thus should offer a menu of contracts for the customers to choose from. We begin by considering a fully separating model, in which the service provider treats the flexible customers differently from the dedicated ones. We shall call this *the basic model* for labeling convenience. The basic mechanism is *ex post* discriminatory with respect to flexibility/taste. Due to the *revelation principle*, we could restrict ourselves to the *direct revelation mechanism* without loss of generality. By such mechanism, we focus on a special class of policies defined in Afeche (2013) as *admissible* policies, which require that policies be stationary, non-anticipative, and independent of arrival processes or service requirements.

The complete contracts consist of the following six options:

$$\left\{ \begin{array}{c} (W_H^1, P_H^1), (W_L^1, P_L^1), (W_H^2, P_H^2), (W_L^2, P_L^2), \\ (W_{Hf}^1, W_{Hf}^2, P_H^f, r_H), (W_{Lf}^1, W_{Lf}^2, P_L^f, r_L) \end{array} \right\},$$

For notational convenience, we define the index sets $T = \{H, L\}$, $M = \{1, 2\}$, $K = \{1, 2, f\}$, $N = \{H, L, Hf, Lf\}$. N is the index set for priority classes *ex post*, where "H" refers to the dedicated H-type customers, "L" refers to the dedicated L-type, and "Hf, Lf" refer to the flexible H-type, L-type respectively. $\{P_i^k\}$'s, $i \in T$, $k \in K$, are the prices charged for the type-i customers who are dedicated to the first, second queue, and flexible customers respectively. r_i represents the probability of the flexible type-i customers being routed to the first queue, whereas with probability $1 - r_i$ they are routed to the second queue. Finally, $\{W_i^k\}$'s, $i \in T$, $k \in K$, are the expected delays for the corresponding segments of customers. By this definition, W_{Hf}^m and W_{Lf}^m are the expected delays for the H-type or the L-type flexible customers given that they are routed to queue $m \in M = \{1, 2\}$. Therefore, the expected delays for the type-i flexible customers could be calculated as follows:

$$W_H^f = r_H W_{Hf}^1 + (1 - r_H) W_{Hf}^2,$$

$$W_L^f = r_L W_{Lf}^1 + (1 - r_L) W_{Lf}^2.$$

Note that an alternative contracts specification could be:

$$\left\{ \begin{array}{c} (W_H^1, P_H^1), (W_H^2, P_H^2), (W_H^f, P_H^f), \\ (W_L^1, P_L^1), (W_L^2, P_L^2), (W_L^f, P_L^f) \end{array} \right\},$$

without specifying the routing probabilities. In terms of tractability, keeping the routing probabilities in the contracts retains the M/M/1 assumptions, while the alternative aggregated approach will lead to M/M/2 settings. In addition, the implementation of M/M/2 based contracts requires dynamic and state-dependent scheduling policy, while the service provider is assumed to commit to the static policies in this chapter. For clear presentation of the analytic results and intuitions, we choose the current approach.

We justify the static policies as follows. (1) Industry practice. For example, Walmart schedules its grocery delivery using batch optimization before midnight. For cases like Instacart and Spoonrocket, they need to submit *planned schedule* (static) before dynamically implementing it. (2) Problem contexts. The service mechanisms, e.g., priorities between classes, the delivery prices etc., are strategic decisions and decided at planning stage. Thus, static policies are not only relevant proxies for real industry practice, but are also consistent with our problem contexts in supporting service provider's strategic decision making.

Utilities. Customers are assumed to be risk-neutral agents who maximize their individual expected utilities. Customers receive a null utility if they walk away. Therefore, an entering type-i customer in the flexibility class $k \in K$, who pretends to be of type-i', will receive the expected utility:

$$u_k(i'|i) = V - C_i W_{i'}^k - P_{i'}^k, \forall i, i' \in T, \forall k \in K. \tag{4.1}$$

Incentive compatibility. The service provider must ensure that customers report their delay sensitivity truthfully. In other words, the following Incentive Compatibility (IC) constraints are required:

$$u_k(i|i) \geq u_k(i'|i), \forall i, i' \in T, \forall k \in K, \tag{4.2}$$

We also need IC constraints to prevent the flexible customers from disguising themselves as the dedicated ones:

$$u_f(i|i) \geq u_m(i|i), \forall i \in T, \forall m \in M. \tag{4.3}$$

In addition, since we assume that the willingness to pay V is sufficiently large, the dedicated customers never pretend to be the flexible ones. For example, if the H-type customers dedicated to the first queue pretend to be flexible, they receive utility $r_H(V - C_H W^1_{Hf}) + (1 - r_H)(0 - C_H W^2_{Hf}) - P^f_H$, which is less than $V - C_H W^1_H - P^1_H$, when V is sufficiently large. In the food delivery application, a meat-lover pretending to be flexible might be offered kale, and thus she has a nonnegative probability of receiving a null payoff. As long as V is large enough, misreporting her dedication in taste is suboptimal.

Individual rationality. The service provider should also ensure that each type of customers receive at least a null payoff. Otherwise, a customer simply walks away. This brings about the following Individual Rationality (IR) constraints:

$$u_k(i|i) \geq 0, \forall i \in T, \forall k \in K. \tag{4.4}$$

Resource constraints. Finally, we shall introduce the Resource Constraints (RE) for the problem, which take into consideration *system stability* and *work conservation*. For notational convenience, we define *effective arrival rate* for each priority class. Using index set N, the effective traffic rates are calculated as follows:

$$\begin{aligned} &A^1_H = \lambda^1_H, \quad A^1_L = \lambda^1_L, \quad A^1_{Hf} = \lambda^f_H r_H, \quad A^1_{Lf} = \lambda^f_L r_L, \\ &A^2_H = \lambda^2_H, \quad A^2_L = \lambda^2_L, \quad A^2_{Hf} = \lambda^f_H (1 - r_H), \quad A^2_{Lf} = \lambda^f_L (1 - r_L). \end{aligned} \tag{4.5}$$

We need System Stability (ST) for all customers in each queue:

$$\sum_{i \in N} A^m_i < \mu, \forall m \in M. \tag{4.6}$$

To highlight the major trade-off and simplify the analysis, throughout the chapter we assume that (ST) should hold at all times, as we focus on the situation where the service capacity is sufficient to cater for all demands.

The *conservation law* should hold for any segment in the power set of N:

$$\sum_{i \in S} \frac{A^m_i W^m_i}{\mu} \geq \frac{\sum_{i \in S} A^m_i / \mu}{\mu - \sum_{i \in S} A^m_i}, \forall m \in M, \forall S \subset N. \tag{4.7}$$

The left-hand side of the conservation law is the expected steady-state remaining service time for the customers of priority classes in the set S due to the Little's law. The right-hand side is the expected waiting time when customers in set S are given absolute priority over all customers outside this set. The constraints follow from classical queueing theory, e.g., Shanthikumar and Yao (1992), while the inequality relaxation is due to Afeche (2004).

The service provider's goal is to maximize the total expected revenue generated from both queues. For a given menu of contracts, her payoff is represented by the objective function as follows:

$$\Pi = P_H^f(A_{Hf}^1 + A_{Hf}^2) + P_L^f(A_{Lf}^1 + A_{Lf}^2) + \sum_{i\in T, m\in M} P_i^m A_i^m. \tag{4.8}$$

To summarize, the optimization problem is formulated as follows:

$$\begin{aligned}
\underset{\{W_i^k, P_i^k, r_i, \forall k\in K, \forall i\in T\}}{\text{Maximize}} \quad & \Pi, \\
\text{subject to} \quad & u_k(i|i) \geq u_k(i'|i), \forall i, i' \in T, \forall k \in K, \\
& u_f(i|i) \geq u_m(i|i), \forall i \in T, \forall m \in M, \\
& u_k(i|i) \geq 0, \forall i \in T, \forall k \in K, \\
& \sum_{i\in N} A_i^m < \mu, \forall m \in M, \\
& \sum_{i\in S} \frac{A_i^m W_i^m}{\mu} \geq \frac{\sum_{i\in S} A_i^m/\mu}{\mu - \sum_{i\in S} A_i^m}, \forall m \in M, \forall S \subset N, \\
& W_i^k, P_i^k \geq 0, 0 \leq r_i \leq 1, \forall k \in K, \forall i \in T.
\end{aligned} \tag{P-1}$$

Numerical analysis

In this subsection, instead of solving the problem analytically, we perform a numerical analysis to get intuitive results. Tables 4.1 and 4.2 display the quantile statistics of the parameters we care about in the basic model. We generate 1000 optimization instances with random traffic λ_L^1, λ_L^f, λ_L^2, λ_H^1, λ_H^f, $\lambda_H^2 \sim U[0, 0.4]$. Valuation $V = 20$, $C_H = 2$, and $C_L = 1$. We normalize μ to be unit service rate.

Table 4.1: Statistics for optimal prices in the basic model.

	P_L^1	P_L^2	P_H^1	P_H^2	P_H^f	P_L^f
75% Quantile	16.19	16.12	17.77	17.77	17.34	16.07
Median	14.99	15.10	17.38	17.40	16.99	15.02
25% Quantile	13.46	13.38	17.00	16.98	16.61	13.15

Table 4.2: Statistics for optimal expected delays in the basic model.

	W_L^1	W_L^2	W_H^1	W_H^2	W_{Lf}^1	W_{Lf}^2	W_{Hf}^1	W_{Hf}^2
75% Quantile	5.12	5.16	1.50	1.51	5.10	4.80	1.96	1.89
Median	3.64	3.57	1.31	1.30	3.42	3.28	1.56	1.52
25% Quantile	2.57	2.64	1.11	1.11	2.47	2.44	1.31	1.30

We notice from the table that the L-type customers are scheduled to wait for significantly longer time than the H-type ones. Consequently, the H-type customers need to pay more to the service provider. This is consistent with the standard results in mechanism design theory. Prices charged for the H-type customers remain roughly the same, since their surplus is always extracted by the service provider. The L-type customers, however, could pretend to be H-type and thus enjoy the information rent, resulting in lower prices charged of them. From the service provider's perspective, he sacrifices the information rent to achieve incentive compatibility so that the mechanism can be implemented.

Somewhat surprisingly, among the H-type customers, the flexible ones are scheduled to wait for longer time than the dedicated one. To understand this observation, we calculate the proportion of the randomly generated instances in which the traffic intensities to both queues are balanced. We find that $P(|\sum_{i\in\{H,Hf\}} A_i^1 - \sum_{i\in\{H,Hf\}} A_i^2| < 0.05) = 65.5\%$, i.e, the effective arrival rates of the impatient customers to both queues are roughly the same in 65.5% of all instances. Intuitively, when the traffic intensities in both queues are more balanced, the flexible H-type customers are less valuable, and therefore suffer from longer delays than the dedicated H-type ones. The common wisdom to assign higher priority to the flexible H-type customers only holds for certain instances in which the traffic intensities in both queues are highly unbalanced.

The value of information

In this subsection, we discuss the value of information in the basic model. We shall show that if the service provider knows each customer's delay sensitivity, he equivalently knows the complete information. On the other hand, however, if the service provider only knows each customer's flexibility, he is not informed of their delay sensitivity for free, and still needs to pay information rent to the L-type customers. Without loss of generality, we normalize μ to be unit service rate.

We begin with a brief illustration of the first-best solution. For given arrival rates and given expected delays, the conditionally optimal pricing schemes are such that the service provider could actually extract the entire surplus from customers:

$$P_i^m = V - C_i W_i^m, \forall m \in M, \forall i \in T. \tag{4.9}$$

The conditionally optimal scheduling policy is directly available through binding RE con-

straints:

$$W_H^m = W_{Hf}^m = \frac{1}{1 - A_H^m}, \forall m \in M. \tag{4.10}$$

$$W_L^m = W_{Lf}^m = \frac{1}{(1 - A_H^m)(1 - A_L^m - A_H^m)}, \forall m \in M. \tag{4.11}$$

In other words, the H-type customers enjoy absolute priority over the L-type customers, regardless of whether they are flexible or not. Since a pooling strategy is optimal, there is no need to differentiate between flexible and dedicated customers *ex post*.

Observable delay sensitivity and unobservable flexibility.

Examples of such information structure include the Internet data service operations. In such systems, customers are routed to different content providers via ISPs (Internet service providers), who charge customers different prices depending on the data speeds and the choice of content providers. Flexible customers could be served by multiple content providers. In this context, the delay sensitivity is usually observable because data speeds depend on technologies via predetermined contract, and the flexibility is customers' private information because they have private preferences over the content providers.

Observation 1: *When customers' delay sensitivity attributes are observable, the first-best is restored.*

In other words, the H-type customers enjoy absolute priority over the L-type customers, regardless of whether they are flexible or not. Intuitively this implies that when the delay sensitivity is observable, the flexible and dedicated customers are equally valuable and enjoy the same scheduling priority. Since a pooling strategy is optimal, there is no need to differentiate between flexible and dedicated customers *ex post*, and the first-best is restored. Intuitively, when the dedicated customers are fully exploited, the flexible customers also receive zero utility by pretending to be dedicated. Since there is no opportunity of gains, the flexible customers will truthfully report their flexibility as well.

Unobservable delay sensitivity and observable flexibility.

This scenario corresponds to the service operations of the electric vehicle charging stations. In this context, the flexibility is observable because it depends on the compatibility of batteries, and the delay sensitivity is drivers' private information.

Observation 2: *If the flexibility is observable and delay sensitivity remains private information, the service provider still needs to pay information rent to the patient customers.*

If we denote $C_H^m = C_H + (C_H - C_L)\frac{A_L^m}{A_H^m}$, $C_{Hf}^m = C_H + (C_H - C_L)\frac{A_{Lf}^m}{A_{Hf}^m}, \forall m \in \{1, 2\}$, as the corresponding *adjusted* waiting cost, the objective function takes the same form as the complete information benchmark. The effect of observable flexibility with unobservable delay sensitivity, is equivalent to an increased gap between the L-type customers and the H-type ones. Therefore, the service provider still needs to pay information rent to the L-type

customers. Under such information structure, the optimal scheduling policy is given in the following proposition:

Proposition 21 *For given arrival rates and pricing schemes, the optimal scheduling policies with unobservable delay sensitivity but observable flexibility share the following characteristics:*

- *The H-type customers (either flexible or dedicated) should be given absolute priority over the L-type customers in both queues.*
- *If $\frac{A_L^m}{A_H^m} > \frac{A_{Lf}^m}{A_{Hf}^m}, \forall m \in M$, then the dedicated H-type customers have the highest priority in queue m. Otherwise, the flexible H-type customers enjoy the highest priority in queue m.*

From the above proposition we know that we need to assign different priorities to the H-type customers depending on their flexibility, while the L-type customers could be treated as the same class. Let W_{AL}^m be the expected steady-state delays for all L-type customers, whether they should be flexible or dedicated. We have the optimal scheduling solutions summarized in Table 4.3.

Table 4.3: Optimal scheduling policy with unobservable delay sensitivity and observable flexibility.

Traffic Regime	$\frac{A_L^m}{A_H^m} > \frac{A_{Lf}^m}{A_{Hf}^m}$	$\frac{A_L^m}{A_H^m} < \frac{A_{Lf}^m}{A_{Hf}^m}$
W_H^m	$\frac{1}{1-A_H^m}$	$\frac{1}{(1-A_{Hf}^m)(1-A_{Hf}^m-A_H^m)}$
W_{Hf}^m	$\frac{1}{(1-A_H^m)(1-A_{Hf}^m-A_H^m)}$	$\frac{1}{1-A_{Hf}^m}$
W_{AL}^m	$\frac{1}{(1-A_H^m-A_{Hf}^m)(1-A_H^m-A_{Hf}^m-A_L^m-A_{Lf}^m)}$	

If all the customers are flexible and the routing decisions are made *ex post*, the model is equivalent to a $M/M/2$ model. If we do not differentiate between the flexible and the dedicated customers, the model degenerates to two separate $M/M/1$ queues with pooling equilibria. The semi-separating equilibrium in our model is due to the information structure that is less studied in the existing literature.

4.4 Server-Specific Mechanism

Model formulation

In the previous discussion, we assume that the service provider treats the flexible customers differently from the dedicated ones *ex post*, and the proposed fully separating mechanism requires six contracts. However, two issues motivate us to consider an alternative mechanism:

1. It is practically difficult to specify routing probabilities in the contracts;

2. If the customers choose which queues to join, it is more natural to consider contract at the server level instead of the system level.

To illustrate the second point, consider the situation where the service provider no longer specifies routing probabilities in the contracts, and thus cannot identify the flexible customers *ex ante*. In this situation, customers decide which queue to join by themselves. In other words, the service provider no longer differentiates between the flexible customers and the dedicated ones *ex post*. Note that customers are heterogeneous only in terms of delay sensitivity but not flexibility. Once they join a particular queue, it is sufficient to consider contracts offered at the server level without loss of generality. Therefore we call this alternative model *server-specific* mechanism.

In this server-specific model, we explore a restricted mechanism space, and assume that pricing and scheduling policies for the flexible customers are identical with the dedicated ones in the queue to which they join. In addition, the model admits further analysis from the marketing perspective as we allow the service provider to decide whether to admit or reject a particular segment of customers. In Proposition 22, we further demonstrate the connections between the basic model and the server-specific model. For clearer presentation of the model, we begin by introducing the complete server-specific contracts.

Contracts. The contracts for the dedicated customers specify the expected delays W_i^m, prices P_i^m and admission controls q_i^m, where $m \in M$, and $i \in T$. In particular, $\{q_i^m\}$'s are the binary decision variables indicating whether a particular segment of customers will be served (q_i^m=1) or rejected (q_i^m=0). The complete menu consists of four contracts:

$$\left\{ \begin{array}{l} (W_H^1, P_H^1, q_H^1), (W_L^1, P_L^1, q_L^1), \\ (W_H^2, P_H^2, q_H^2), (W_L^2, P_L^2, q_L^2) \end{array} \right\}.$$

Sequence of events. The service provider offers take-it-or-leave-it contracts for arriving customers. Informed of the contracts specifications, customers decide whether they should join or balk. In addition, the flexible customers also decide which queue to join. Once the dedicated customers enter, they receive service if they are accepted, wait for the specified expected delay, and make payment. Once the flexible customers enter, they could be served in the chosen queue through the same service procedure as the dedicated customers in that queue.

Utilities. We start with the micro-structure of customers' decision-making process. A dedicated type-i customer arrives at the queue m, who pretends to be of type-i', will receive the expected utility:

$$u_m(i'|i) = q_{i'}^m(V - C_i W_{i'}^m - P_{i'}^m), \forall m \in M, \forall i, i' \in T. \tag{4.12}$$

On the other hand, we define $\{r_i\}$'s as the probabilities of the flexible customers joining the first queue. We call the $\{r_i\}$'s the *self-adaptive* routing probabilities. We assume that

$r_i \in (0,1)$, $\forall i \in T$. Thus, a flexible customer of type-i pretending to be type-i' receives the expected utility:

$$\begin{aligned} u_f(i'|i) &= q^1_{i'} r_{i'}(V - C_i W^1_{i'} - P^1_{i'}) + q^2_{i'}(1 - r_{i'})(V - C_i W^2_{i'} - P^2_{i'}) \\ &= u_1(i'|i) r_{i'} + u_2(i'|i)(1 - r_{i'}). \end{aligned} \tag{4.13}$$

The fact that the self-adaptive routing probabilities are not corner solutions implies that each flexible customer plays a mixed strategy and randomizes between joining two queues. We only analyze such mixed strategies to highlight major trade-offs.

Individual rationality. The service provider should also ensure that each type of customers receive at least a null payoff. Otherwise, a customer could simply walk away. This brings about the following (IR) constraints:

$$u_k(i|i) \geq 0, \forall i \in T, \forall k \in K. \tag{4.14}$$

Since the utility for a flexible customer is the convex combination of that for dedicated customers to the two queues respectively, the IR constraints for the flexible customers are redundant.

Indifference decision. We introduce the Indifference Decision (ID) constraints:

$$u_1(i|i) = u_2(i|i) = u_f(i|i), \forall i \in T. \tag{4.15}$$

In equilibrium, in order to induce a flexible customer to randomize over two queues, she must feel indifferent between them. Otherwise, she never chooses the server that gives her a strictly lower expected utility. If flexible customers are not indifferent between joining either queue, they will not truthfully report their flexibility, as they must be better off by pretending to be some dedicated customers. Notice that we implicitly assume that the dedicated customers cannot pretend to be the flexible ones, for the same reason as in the basic model: The null utility for not receiving service naturally penalizes customers from misreporting their dedication, as long as V is sufficiently large.

Resource constraints. As in the basic model, we define the total *effective arrival rate* of type-i customers to the first queue as:

$$A^1_i = \lambda^1_i q^1_i + \lambda^f_i q^1_i r_i, \forall i \in T, \tag{4.16}$$

and similarly for the second queue:

$$A^2_i = \lambda^2_i q^2_i + \lambda^f_i q^2_i (1 - r_i), \forall i \in T, \tag{4.17}$$

which characterizes the traffic intensities in the corresponding queue. Using such notations, the system stability for queue m requires:

$$A^m_H + A^m_L < \mu, \forall m \in M. \tag{4.18}$$

The conservation law for queue m requires:

$$\sum_{i\in S}\frac{A_i^m W_i^m}{\mu}\geq\frac{\sum_{i\in S}A_i^m/\mu}{\mu-\sum_{i\in S}A_i^m},\forall m\in M,\forall S\subset T. \tag{4.19}$$

For a given menu of contracts, the service provider's total expected revenue generated from both queues is as follows:

$$\Pi=\sum_{m\in M}\sum_{i\in T}A_i^m P_i^m. \tag{4.20}$$

Now we can piece together the objective function and the constraints in the following formulation:

$$\begin{aligned}
\underset{\{W_i^m,P_i^m,q_i^m,\forall m\in M,\forall i\in T\}}{\text{Maximize}}\ \Pi=&\sum_{m\in M}\sum_{i\in T}A_i^m P_i^m,\\
\text{subject to }& u_k(i|i)\geq u_k(i'|i),\forall i,i'\in T,\forall k\in K,\\
&u_1(i|i)=u_2(i|i),\forall i\in T,\\
&u_k(i|i)\geq 0,\forall i\in T,\forall k\in K,\\
&\sum_{i\in T}A_i^m<\mu,\forall m\in M,\\
&\sum_{i\in S}\frac{A_i^m W_i^m}{\mu}\geq\frac{\sum_{i\in S}A_i^m/\mu}{\mu-\sum_{i\in S}A_i^m},\forall m\in M,\forall S\subset T,\\
&W_i^m,P_i^m\geq 0,q_i^m\in\{0,1\},r_i\in(0,1),\forall m\in M,\forall i\in T.
\end{aligned} \tag{P-2}$$

Comparing with the optimization problem $(P-1)$, we notice that there are major differences in $(P-2)$. First, the objective function consists of four segments instead of six. Second, there is an additional (ID) constraint. Third, there are only two priority classes in each queue instead of four. Before solving the model, we build a connection between the basic model and the server-specific model.

Proposition 22 *The server-specific mechanism is equivalent to a particular separating mechanism, in which the service provider offers the following contracts:*

$$\left\{\begin{array}{c}(W_H^1,P_H^1,q_H^1),(W_L^1,P_L^1,q_L^1),(W_H^2,P_H^2,q_H^2),(W_L^2,P_L^2,q_L^2),\\(W_H^1,P_H^1,q_H^1,W_H^2,P_H^2,q_H^2,r_H),(W_L^1,P_L^1,q_L^1,W_L^2,P_L^2,q_L^2,r_L)\end{array}\right\}.$$

From the proposition we can see, when all customers are admitted, the server-specific mechanism restricts the contracts specifications W_i^f as convex combinations of W_i^1 and W_i^2, which is a special case of the basic model, where $W_H^f=r_H W_{Hf}^1+(1-r_H)W_{Hf}^2$, $W_L^f=r_L W_{Lf}^1+(1-r_L)W_{Lf}^2$. Therefore the server-specific mechanism does not make full use of the discrimination instrument, and leads to less revenue due to such restrictions.

Our analysis for the server-specific model consists of two stages.

- **Stage one**: For given arrival rates A_i^m (traffic intensities), i.e., with fixed r_i and q_i^m, determine the optimal expected delays W_i^m and prices P_i^m, for $\forall m \in M, \forall i \in T$.

- **Stage two**: Determine the optimal effective arrival rates A_i^m, $\forall m \in M, \forall i \in T$, by optimizing over joining probabilities $0 < r_i < 1$ and choosing binary controls q_i^m, $\forall i \in T$, while W_i^m, P_i^m are jointly optimized for $\forall m \in M, \forall i \in T$.

Note that it is equivalent for the service provider to optimize over r_i in the second stage, even though the customers self-select which queue to join. This equivalence is due to the (ID) constraints, which establish a one-to-one mapping from $\{W_i^m, P_i^m, q_i^m, \forall m \in M\}$ to r_i, $\forall i \in T$. In what follows, we solve the first-stage problem for given effective arrival rates. We first decompose the master problem into four cases by identifying some structural properties of the admission control policies. For each case, we work out the conditionally optimal pricing and scheduling policies given fixed admission controls. Then we discuss the second-stage problem concerning the customers' equilibrium queue-joining choices.

Conditionally optimal scheduling policies

We restrict ourselves to binary admission control policies, i.e., $q_i^m \in \{0, 1\}$, $\forall m \in M, \forall i \in T$, categorizing the solutions by different admission regimes.

Lemma 4 *A server will not serve only customers with high delay sensitivity. Furthermore, the IC constraints for the flexible customers are redundant.*

As direct consequences of the above lemma, there are four possible solution regimes to discuss:

- Case 1: $q_L^1 = q_L^2 = q_H^1 = q_H^2 = 1$. Both servers accept both types of customers. We denote this as LH-LH case.

- Case 2: $q_L^1 = q_L^2 = 1, q_H^1 = q_H^2 = 0$. Both servers accept only low delay sensitive customers. We denote this as L-L case.

- Case 3: $q_L^1 = q_L^2 = q_H^1 = 1, q_H^2 = 0$. The first server accepts both types whereas the second server only accepts L-type customers. We denote this as LH-L case.

- Case 4: $q_L^1 = q_L^2 = q_H^2 = 1, q_H^1 = 0$. The second server accepts both types whereas the first server only accepts L-type customers. We denote this as L-LH case.

We analyze the LH-LH case here, while the analysis and results for the other cases are summarized in the appendix.

We now characterize those policies following the "achievable region approach". In line with the definitions in this stream of literature, the delay profiles under *absolute preemptive priority* for either H-type or L-type customers correspond to the solutions where the resource constraint for the particular customer segment is binding. *Work-conserving* refers

to the policies such that servers never idle when there are customers in the queue. The delay profiles under *randomized static preemptive priority* correspond to the Pareto-optimal solutions in terms of the expected delays for both service classes, but no resource constraint for a particular customer segment is binding. If a delay profile deviates from Pareto-optimal solutions, we say there is *inserted job idleness* or *strategic delay*. Without loss of generality, we normalize μ to be unit service rate.

Given specific scheduling policy, the optimization problem of solving for conditionally optimal pricing schemes is a linear programming problem. To solve for the optimal prices, we need the following lemma.

Lemma 5 *Let the optimal solution for the service provider's problem be* (P_i^m, W_i^m, r_i), $\forall i \in T$ *and* $\forall m \in M$. *The following properties hold:*

- *The utility surplus of the L-type customers is strictly greater than that of the H-type customers:*
$$u_m(L|L) > u_m(H|H), \forall m \in M; \tag{4.21}$$
- *The IR constraints for the H-type customers are binding:*
$$u_1(H|H) = u_2(H|H) = 0; \tag{4.22}$$
- *At least one of the IC constraints for the L-type customers is binding:*
$$u_1(L|L) \geq u_1(H|L); \tag{4.23}$$
$$u_2(L|L) \geq u_2(H|L). \tag{4.24}$$

This lemma enables us to identify the pricing strategies. It turns out that under the optimal prices, the H-type customers have zero surplus while the L-type customers receive positive surplus $(C_H - C_L)W_H^1$ or $(C_H - C_L)W_H^2$, which is the information rent due to the IC constraints. Once we identify the pricing strategies (as functions of the expected delays), we can break down the master problem into two subproblems. These two subproblems differ in the binding constraints and therefore differ in the expressions of payments as well as the feasibility conditions:

- Subproblem 1:

$$\begin{aligned}
\underset{\{W_i^m, \forall m \in M, \forall i \in T\}}{\text{Maximize}} \ \Pi &= \sum_{m \in M} \sum_{i \in T} A_i^m P_i^m, \\
\text{subject to } & P_H^m = V - C_H W_H^m, \forall m \in \{1, 2\}, \\
& P_L^1 = V - (C_H - C_L)W_H^1 - C_L W_L^1, \\
& P_L^2 = V - (C_H - C_L)W_H^1 - C_L W_L^2, \\
& W_H^2 \leq W_H^1 \leq W_L^2, W_H^1 \leq W_L^1, \\
& W_i^m \geq 0, (ST), (RE), \forall m \in M, \forall i \in T.
\end{aligned}$$

- Subproblem 2:

$$\begin{aligned}
\underset{\{W_i^m,\forall m\in M,\forall i\in T\}}{\text{Maximize}}\ \Pi &= \sum_{m\in M}\sum_{i\in T} A_i^m P_i^m,\\
\text{subject to } & P_H^m = V - C_H W_H^m, \forall m \in \{1,2\},\\
& P_L^1 = V - (C_H - C_L)W_H^2 - C_L W_L^1,\\
& P_L^2 = V - (C_H - C_L)W_H^2 - C_L W_L^2,\\
& W_H^1 \le W_H^2 \le W_L^1, W_H^2 \le W_L^2,\\
& W_i^m \ge 0, (ST), (RE), \forall m \in M, \forall i \in T.
\end{aligned}$$

Notice that effective arrival rates are functions of the self-adaptive routing probabilities (r_H, r_L), and are assumed given at this stage. Next, we show that the first subproblem can be solved in closed forms, which yield the optimal scheduling policy summarized in Table C.1 in the appendix. Symmetrically we derive the optimal scheduling policy for the service provider in the second subproblem without going through similar analysis, and the results are summarized in Table C.2 in the appendix.

Next, we can compare the two basic feasible solutions of pricing schemes and the results are summarized in the following proposition:

Proposition 23 *For the LH-LH case, given traffic assignments, the optimal expected steady state delays as well as the corresponding scheduling policies in both queues are summarized in Table 4.4.*

When more H-type customers enter the second queue than all customers to the first queue (H-type and L-type combined), there are incentives for the service provider to strategically delay the L-type customers awaiting in the first queue. From the service provider's perspective, he inserts strategic idleness in the first queue with the anticipation that there will soon be the new arrival of H-type customers. If he admits an L-type customer without inserting idleness, the potential H-type arrival needs to wait in the queue and thus incur a much higher cost. From the customers' perspective, such idleness prevents H-type customers from pretending to be L-type ones since the price charged for L-type is much cheaper. This leads to the conditionally optimal solution in the case 1.

When not as many H-type customers enter the second queue, but they still significantly outweigh the inflow to the first queue, a similar phenomenon could happen if the waiting cost ratio $\frac{C_H}{C_L}$ is above certain threshold. This indicates that H-type customers are valuable and the revenue loss from delaying the L-type customers is comparatively less important. This explains the solution in case 2. If the ratio $\frac{C_H}{C_L}$ is below the threshold, or if H-type customers routed to the second queue are less than total customers to the first queue, no strategic delay policy is optimal; this leads to the situation in cases 3 and 4. Cases 5 to 8 are symmetric and their intuitions are similar to cases 1 to 4. The geometric representations of the eight traffic regimes are illustrated as in Figure 4.2.

The results for L-L case and LH-L case are summarized in the appendix. The major intuition remains the same.

Table 4.4: Conditionally optimal scheduling policy for the master problem in the server-specific model.

Case	Regime	Pricing Schemes	Expected Steady State Delays	Scheduling Policies
1	$\frac{1}{1-A_H^2} > \frac{1}{(1-A_H^1)(1-A_H^1-A_L^1)}$	subproblem 2	$W_L^1 = W_H^2 = \frac{1}{1-A_H^2}, W_H^1 = \frac{1}{1-A_H^1}$ $W_L^2 = \frac{1}{(1-A_H^2)(1-A_H^2-A_L^2)}$	Strategic delay in queue 1
2	$\frac{1}{1-A_H^1-A_L^1} \le \frac{1}{1-A_H^2} \le \frac{1}{(1-A_H^1)(1-A_H^1-A_L^1)}$ $\frac{C_H}{C_L} \ge \frac{A_L^1[(A_H^2-A_L^1-A_H^1)-A_H^1(1-A_H^1-A_L^1)]}{A_H^1(A_H^1-A_H^2)(1-A_H^1-A_H^2)}$	subproblem 1	$W_L^1 = W_H^1 = W_H^2 = \frac{1}{1-A_H^2}$ $W_L^2 = \frac{1}{(1-A_H^2)(1-A_H^2-A_L^2)}$	Absolute preemptive priority in queue 2
3	$\frac{1}{1-A_H^1-A_L^1} \le \frac{1}{1-A_H^2} \le \frac{1}{(1-A_H^1)(1-A_H^1-A_L^1)}$ $\frac{C_H}{C_L} \le \frac{A_L^1[(A_H^2-A_L^1-A_H^1)-A_H^1(1-A_H^1-A_L^1)]}{A_H^1(A_H^1-A_H^2)(1-A_H^1-A_H^2)}$	subproblem 2		
4	$\frac{1}{1-A_H^1} \le \frac{1}{1-A_H^2} \le \frac{1}{1-A_H^1-A_L^1}$		$W_H^1 = \frac{1}{1-A_H^1}, W_L^1 = \frac{1}{(1-A_H^1)(1-A_H^1-A_L^1)}$	Absolute preemptive priority in queue 1
5	$\frac{1}{1-A_H^2} \le \frac{1}{1-A_H^1} \le \frac{1}{1-A_H^2-A_L^2}$	subproblem 1	$W_H^2 = \frac{1}{1-A_H^2}, W_L^2 = \frac{1}{(1-A_H^2)(1-A_H^2-A_L^2)}$	Absolute preemptive priority in queue 2
6	$\frac{1}{1-A_H^2-A_L^2} \le \frac{1}{1-A_H^1} \le \frac{1}{(1-A_H^2)(1-A_H^2-A_L^2)}$ $\frac{C_H}{C_L} \le \frac{A_L^2[(A_H^1-A_L^2-A_H^2)-A_H^2(1-A_H^2-A_L^2)]}{A_H^2(A_H^2-A_H^1)(1-A_H^1-A_H^2)}$			
7	$\frac{1}{1-A_H^2-A_L^2} \le \frac{1}{1-A_H^1} \le \frac{1}{(1-A_H^2)(1-A_H^2-A_L^2)}$ $\frac{C_H}{C_L} \ge \frac{A_L^2[(A_H^1-A_L^2-A_H^2)-A_H^2(1-A_H^2-A_L^2)]}{A_H^2(A_H^2-A_H^1)(1-A_H^1-A_H^2)}$	subproblem 2	$W_L^1 = \frac{1}{(1-A_H^1)(1-A_H^1-A_L^1)}$ $W_L^2 = W_H^2 = W_H^1 = \frac{1}{1-A_H^1}$	Absolute preemptive priority in queue 1
8	$\frac{1}{1-A_H^1} > \frac{1}{(1-A_H^2)(1-A_H^2-A_L^2)}$	subproblem 1	$W_L^1 = \frac{1}{(1-A_H^1)(1-A_H^1-A_L^1)}$ $W_L^2 = W_H^1 = \frac{1}{1-A_H^1}, W_H^2 = \frac{1}{1-A_H^2}$	Strategic delay in queue 2

Self-adaptive routing

We have so far been able to design the optimal mechanism for given arrival rates. Recall that, in the server-specific model, the flexible customers also decide which queue to join. Therefore, "self-adaptive routing" indicates that customers self-select which queue to join and choose a mixed joining strategy. In this subsection, we characterize the mixed-strategy routing equilibrium for a given admission control policy. First, we shall solve the routing problem in each case listed in Table 4.4, for which we need additional sufficient conditions to achieve closed-form results.

Proposition 24 *In the LH-LH scenario, the self-adaptive routing probabilities in each of the eight cases are give as follows:*

- *In case* 1*, when* $\lambda_H^2 > \lambda_H^1 + \lambda_H^f + \lambda_L^1 + \lambda_L^f + (\lambda_H^1 + \lambda_H^f)(1 - \lambda_H^1 - \lambda_H^f - \lambda_L^1 - \lambda_L^f)$*, equilibrium routing probabilities* $r_H, r_L \to 1$*;*
- *In case* 2*, when* $\lambda_H^1 + \lambda_H^f + \lambda_L^1 + \lambda_L^f \le \lambda_H^2 \le \lambda_H^1 + \lambda_H^f + \lambda_L^1 + \lambda_L^f + (\lambda_H^1 + \lambda_H^f)(1 - \lambda_H^1 - \lambda_H^f - \lambda_L^1 - \lambda_L^f)$*, and* $\frac{C_H}{C_L} \ge \frac{1-A_L^2-(1-A_H^2)^2/(1-A_H^2-A_L^2)^2}{A_H^1+A_L^1+A_H^2+A_L^2}$*, equilibrium routing probabilities*

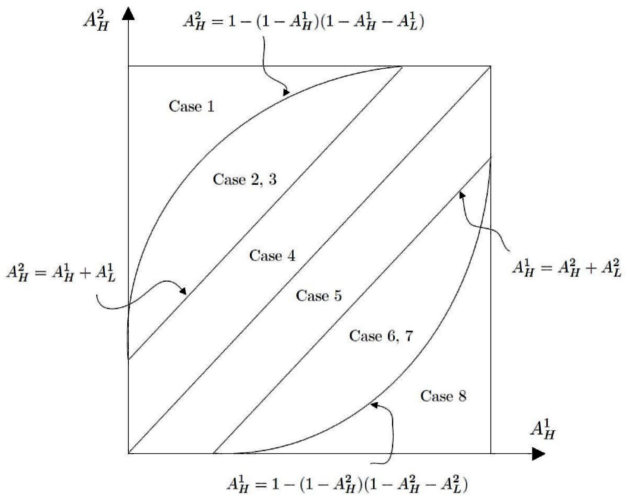

Figure 4.2: Geometric representation of the eight cases in LH-LH scenario.

$r_H, r_L \to 1$;

- *In case* 3 *and case* 4, *when* $\lambda_H^f < \lambda_H^2 - \lambda_H^1$, *if* $\lambda_L^f \geq |\lambda_H^2 + \lambda_L^2 - \lambda_H^1 - \lambda_L^1 - \lambda_H^f|$, *equilibrium routing probabilities are* $r_H \to 1$, $r_L = \frac{1}{2} + \frac{\lambda_H^2+\lambda_L^2-\lambda_H^1-\lambda_L^1-\lambda_H^f}{2\lambda_L^f}$, *resulting in* $A_H^1 + A_L^1 = A_H^2 + A_L^2$;
- *Special case: in case* 4, *if* $\lambda_L^f \geq |\lambda_L^2 - \lambda_L^1|$ *and* $\lambda_H^f \geq |\lambda_H^2 - \lambda_H^1|$, *equilibrium routing probabilities are* $r_H = \frac{1}{2} + \frac{\lambda_H^2-\lambda_H^1}{2\lambda_H^f}$, $r_L = \frac{1}{2} + \frac{\lambda_L^2-\lambda_L^1}{2\lambda_L^f}$, *resulting in* $A_H^2 = A_H^1$;
- *Cases* 5, 6, 7, 8 *are symmetric cases corresponding to cases* 4, 3, 2, 1 *respectively.*

The results suggest that the H-type flexible customers choose a mixed-strategy so that their arrival rates to both queues are the same whenever possible. The L-type flexible customers choose a mixed-strategy so that total arrival rates to both queues are the same whenever possible. If the arrival rates to both queues are well-balanced, it leads to the special case. If (almost) all the flexible H-type customers choose to join the first queue, the L-type customers choose a mixed strategy so that the combined traffic (both H-type and L-type) is balanced between the two queues. This leads to cases 3 and 4. In the most extreme situations, (almost) all flexible customers join the first queue, which are cases 1 and 2.

The results shed interesting light on the connection between the server-specific model and the basic model. In the server-specific model, the flexible customers' equilibrium queue-joining choices are *as if* the service provider is implementing a *hierarchical load-balancing* routing algorithm in the basic model: The service provider should first route the flexible H-type customers to balance the arrival rates of the H-type customers to both queues, and then route the flexible L-type customers so that total arrival rates to both queues are the

same whenever possible. As long as the sufficient conditions in Proposition 24 are satisfied, the solutions by the *hierarchical load-balancing* routing policy are exact for the *LH-LH* scenario. Otherwise, this algorithm gives an approximation of the optimal solution.

We provide numerical analysis to evaluate the performance of the *hierarchical load-balancing* policy as a potential heuristic algorithm in solving the second-stage problem. We generate 10000 scenarios with random traffic inputs λ_L^1, λ_L^f, λ_L^2, λ_H^1, λ_H^f, $\lambda_H^2 \sim U[0, 0.4]$ and unit service rate. Valuation $V = 30$, $C_H = 2$, and $C_L = 1$. In Table 4.5, the proportion of each traffic regime is summarized according to the eight cases categorized as in Table 4.4. We calibrate $P(|r_i - r_i^{baseline}| < 0.05)$, $\forall i \in T$, as the probabilities that the proposed heuristic algorithm yields solutions sufficiently close to the baselines (the solutions given by nonlinear optimization). For cases 1, 2, 7, 8 and the "special case" in Table 4.4, the algorithm yields

Table 4.5: Algorithm performance in the *LH-LH* case.

	Case 1	Cases 2, 3	Case 4	Special Case	Case 5	Cases 6,7	Case 8
Number of Instances	32	71	2483	4824	2484	66	40
$P(\lvert r_H - r_H^{baseline}\rvert < 0.05)$	1.000	0.848	0.902	1.000	0.726	0.855	1.000
$P(\lvert r_L - r_L^{baseline}\rvert < 0.05)$	1.000	0.894	0.999	1.000	1.000	0.887	1.000

the same solutions as nonlinear optimization, while in other cases, there are gaps in the performance. In addition, the same method could be applied to the *L-L* and *LH-L* scenarios, for which we are guaranteed of optimal solutions under some sufficient conditions. We also summarize the results in the appendix.

Optimal admission control

Since we adopt binary admission control policies, it suffices to make a static comparison among different admission policies to find the optimal one. We shall next restrict ourselves to two representative traffic regimes serving as the common ground for comparison. The next proposition captures the major trade-off in the admission control; similar results apply to other regimes as well.

- *Balanced traffic regime*, where $\lambda_L^f \geq |\lambda_L^2 - \lambda_L^1|$ and $\lambda_H^f \geq |\lambda_H^2 - \lambda_H^1|$, and $\lambda_L^f \geq |\lambda_L^2 - \lambda_H^1 - \lambda_L^1 - \lambda_H^f|$;
- *Unbalanced traffic regime*, where $\lambda_H^1 > 1 - (1 - \lambda_H^2 - \lambda_H^f)(1 - \lambda_H^2 - \lambda_H^f - \lambda_L^2 - \lambda_L^f)$, and $\lambda_L^f > |\lambda_L^1 - \lambda_L^2|$.

Proposition 25 *In either the balanced traffic regime or the unbalanced traffic regime, there exist valuation thresholds $\overline{V}$ and $\underline{V}$, such that if $V \geq \overline{V}$, it is optimal to choose LH-LH policy; if $V \leq \underline{V}$, we should choose L-L policy; otherwise if $\underline{V} \leq V \leq \overline{V}$, LH-L policy is optimal among the three.*

The economic intuitions behind this proposition are clear: If we shut down the admission channel for a particular priority class, we lose potential revenue generated by this segment of customers. In contrast, admitting too many L-type customers might result in traffic congestion and jeopardize the additional value that could have been generated by serving H-type customers.

Model comparison

First, we compare the service provider's revenue in the basic model with that in the server-specific model to get some sense of the scale of the revenue loss. In Figure 4.3, given unit service rate $\mu = 1$, we generate fixed arrival rates $\lambda_L^1 = 0.1900$, $\lambda_L^f = 0.0462$, $\lambda_L^2 = 0.1214$, $\lambda_H^1 = 0.0972$, $\lambda_H^f = 0.1783$, $\lambda_H^2 = 0.3048$. Valuation $V = 20$ is sufficiently high, and $C_H = 2$, $C_L = 1$. We generate 100 numerical instances, for which we randomly generate routing probabilities r_L,$r_H \sim U[0, 1]$.

For each simulated instance, since the routing probabilities are given, and the valuation V is large enough to admit all customer segments in the server-specific model, we can calibrate the impact of *ex post* discrimination alone. Notice that the revenue generated by the discriminatory policies in the basic model dominates that in the server-specific model over all realization of routing randomness. However, the revenue loss is less than 1% of total revenue, while the impact of routing consists of around 5% of total revenue. By randomizing over the traffic inputs, we check that this scale is robust. This means that, the server-specific model will suffice in many practical applications.

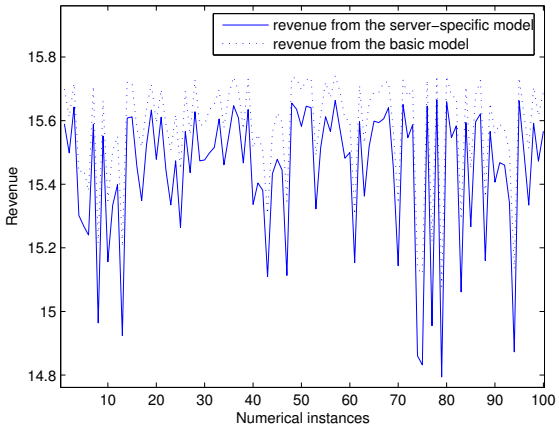

Figure 4.3: Revenue comparison with coupled routing probabilities.

By comparing between the basic model and the server-specific model, we can also gain some managerial insights on the impact of the *ex post* discrimination with respect to customers' flexibility. If the flexible and impatient customers suffer from longer delay than

the dedicated ones, then the dedicated impatient customers enjoy shorter delay in the basic model than in the server-specific model. We show this by numerical analysis with the coupled arrival rates. We generate 1000 instances with random traffic λ_L^1, λ_L^f, λ_L^2, λ_H^1, λ_H^f, $\lambda_H^2 \sim U[0, 0.4]$. Valuation $V = 20$, $C_H = 2$, and $C_L = 1$. It turns out that the median expected delays are $W_H^1 = 1.44$, $W_H^2 = 1.45$ in the server-specific model, while the corresponding median dropped to $W_H^1 = 1.31$, $W_H^2 = 1.30$ in the basic model, and the median for the flexible customers increased to $W_{Hf}^1 = 1.56$, $W_{Hf}^2 = 1.52$. Akgun et al. (2012) prove that the dedicated customers always benefit from the presence of flexible customers, because they assume perfect information. Our model suggests that the dedicated customers could be deprioritized with the increasing fraction of the flexible customers under unbalanced traffic, and then the impact of flexible customers becomes positive when the traffic intensities are balanced.

4.5 Conclusion

In this chapter, we discuss the revenue maximization problem in service operations systems where customers are heterogeneous both in terms of the delay sensitivity (patience) and taste preference (flexibility). In the basic model, we investigate a horizontal substitution strategy which incorporates *ex post* discrimination between the dedicated customers and the flexible ones after they are admitted for service. We solve for the jointly optimal pricing strategies, steady-state scheduling rules, probabilistic routing policies. Compared with the dedicated customers, the flexible ones enjoy further information rent since their flexibility is also private information. Furthermore, the first-best (defined in Section 4.3) could be restored if the delay sensitivity is observable, while the service provider should still pay information rent to the patient customers when only the flexibility is observable.

Realizing that the horizontal substitution strategy in the basic model requires specification of routing probabilities for the flexible customers, which might cause implementation difficulties in practice, we propose an alternative *server-specific* model wherein the flexible customers self-select which queue to join. Once a flexible customer joins any of the two queues, she is identical with the dedicated customers in terms of service priorities. We find that the service provider loses some of his discrimination power in the server-specific mechanism, compared with the fully separating basic mechanism. If the customer arrival rates to both queues are relatively balanced, the "$c\mu$ rule" should be adopted and the impatient customers should be assigned absolute preemptive priorities over the patient ones. However, when we have excessive impatient customers in one queue, it is likely that the service provider should insert strategic delays for the patient customers in the other queue due to incentive compatibility. We show by numerical examples that the revenue gap between the basic model and the server-specific model is small, while the latter mechanism is easier to implement.

Chapter 5

Learning with Projection Effects in Service Operations Systems

5.1 Introduction

Service operations systems, e.g., call centers, hospitals, restaurants, and etc., are universal in the modern society. When the customers decide whether they want to join the system and enjoy the service, their decisions depend on the service quality, which is usually unknown *a priori*. In such an environment, a long queue signals both high quality and severe congestion. What still remains a puzzle is that we often observe excessively long queue for low quality service. This is particularly true for the tourism industry. Tripadvisor.com is an online consumer review platform for tourist attractions. Here is a typical review for a popular restaurant in San Francisco[1]:
After seeing the huge queue outside (the restaurant), there was no doubt the place was extremely popular, so we thought to check it out. After a long wait in line, the food was pretty disappointing. There was a hair in the French toast and the sausage was pretty bad. The rest was average at best. Nothing special and way too much hype. Wouldn't return.
How should we rationalize such phenomenon, if we believe that queue lengths are useful instruments to signal quality? In the Tripadvisor example, a simple explanation relies on the heterogeneity of customers: Tourists are usually less time sensitive, while the local customers are relatively impatient. If a tourist over-estimates the population of the local customers, she is happy to wait in the long queue due to over-optimistic expectations of service quality. She fails to realize that the long queue in a tourist attraction may not be informative of the service quality, simply because the majority of the awaiting customers are also tourists who have plenty of leisure time!

The phenomenon is not unique in the tourism industry. ZocDoc.com is a platform on which the patients could read feedback reviews for doctors and make appointments completely online. Similar stories featuring "long wait for bad service" are not uncommon: A

[1] A redacted version taken from tripadvisor.com (retrieved on 2014/12/12).

dentist who specializes in wisdom teeth removal and has been booked full for the next two months may not be worth the wait. The popularity is misinterpreted as an indicator for high quality because most of the patients may not be in urgent conditions, and their appointments in advance only exaggerate the dentist's popularity.

The key ingredients in the examples of Tripadvisor and ZocDoc are the customers' two-dimensional heterogeneities, both in terms of the delay sensitivity (*patience*) and information precision about the service quality (*private signals*). The most notable examples of differentiated delay sensitivity include Amazon's regular delivery and two-day fast delivery. ISPs (Internet service providers) such as AT&T U-verse offer connection speeds ranging from 3 Mbps (Megabits Per Second) to 45 Mbps. On the other hand, examples of the private signals of quality judgement include *word of mouth* or *anecdotes* from their friends or online platforms such as Yelp.com.

The traditional approach in the economics of queues requires accurate estimations of the population information, e.g., the fraction of impatient customers. Although this might be true for the service provider, it is nonetheless a strong assumption for the customers. The wisdom from the examples of Tripadvisor and ZocDoc suggests that, customers usually suffer from bounded rationalities of estimations that are related to their own attributes. It is well-known in psychology that, people often suffer from the "false-consensus effects": They believe that others are similar to themselves, e.g., Ross et al. (1977). The opposite direction of such egocentric bias is called "psychological marginality effects", e.g., Frable (1993). For the convenience of terminology, we follow the economics literature Madarász (2011) and call such bounded rationalities as the "projection effects" and the "reversed-projection effects".

In this chapter, we study how bounded rationality influences learning and subsequently queue-joining behavior. We propose a stylized single-server queueing model with observable queue length to study the service system performances. The customers are endowed with the aforementioned two-dimensional heterogeneity. In the fully rational benchmark, by characterizing the customers' equilibrium queue-joining strategies, we rationalize the mental struggles behind the *hesitation* behaviors, a class of non-monotone strategies under which the uninformed and impatient customers form consensus to stop joining the queue at multiple "holes". By delaying the joining decision until the queue is even longer, an uninformed customer anticipates that the impatient customers would not wait in such congested environment and the remaining customers should be better indicators for the service quality.

Ironically, under projection bias, customers who are more averse to waiting will react more sensitively to the observed long queue. This leads to over-estimation of the service quality and induces them to wait in the long queue. Conversely, under reversed-projection bias, the patient customers tend to under-estimate other customers' patience, over-estimate the service quality, and wait in the long queue. While the reversed-projection effects seem to select the right customers to wait in the queue, they give rise to a different source of inefficiency: The impatient customers with the reversed-projection bias adopt a suboptimal threshold queue-joining strategy, because they learn nothing from the patient customers' behaviors. In other words, the queueing dynamics become uninformative for the impatient customers.

In terms of the social welfare, the bounded rationalities impede effective social learning by inducing type-I decision errors, i.e., balking from the queue when the service quality is high, and type-II errors, i.e., joining the queue when the quality is low. In particular, the (reversed-) projection effects reduce the type-I welfare loss but increase the type-II welfare loss due to the impatient (patient) customers, compared with the fully rational benchmark. Thus, the inefficiency of social learning driven by such bounded rationalities could potentially impair social welfare due to the blind "buying frenzy" even if the service quality is low.

To further evaluate the system performances, we adopt the stochastic comparison approach for the limiting queue lengths under different scenarios. We find that queue lengths are the longest when the impatient customers suffer from the projection effects while the patient customers suffer from the reversed-projection effects. Intuitively, both the patient and the impatient customers under-estimate others' patience, and are simultaneously trapped in the blind "buying frenzy" situations.

The rest of this chapter is organized as follows. Section 5.2 reviews relevant literature. Section 5.3 introduces our model setup. In Section 5.4, we carry out the analysis, which is followed by numerical examples in Section 5.5. Section 5.6 extends the model and evaluates system performances in different models. Section 5.7 concludes. All proofs are provided in the appendix.

5.2 Literature Review

Our model is related to the game-theoretic social learning literature in economics, which incorporates both information and incentive externalities. Dasgupta et al. (2000) discusses a sequential investment game, in which a "balking" move incurs an epic loss due to negative payoff complementarities. He characterizes the *trigger equilibria* that depend on the agents' private signals, and shows that there would be *strong herding* for signals with bounded support. Zhou and Chen (2015) offer an alternative model of the sequential decision problem by network signaling, but they focus on the case with positive incentive externalities. Eyster et al. (2014) study observational learning with negative payoff complementarities by *congestions*. Under their settings, complete learning may also fail when agents herd. We deviate from their regime by assuming that, the congestion costs would eventually become sufficiently high with the increasing population size of one's predecessors, thereby breaking the uninformative herd.

Our work is also in line with the service operations literature. Veeraraghavan and Debo (2009) consider a model of two queues, in which customers receive imperfect binary signals about the unobservable service quality. They rationalize a customer's behavior to ignore her private information and "join the longer queue". The signaling effects strengthen if the unknown service rates are negatively correlated with the service quality, and vise versa if the correlation is positive. Veeraraghavan and Debo (2011) study a similar model with only one queue, and incorporate congestion costs in the customers' utility functions. Debo and Veeraraghavan (2014) extend the previous single-server queueing model such that both the

service rate and the service quality are unobservable *ex ante* but are positively correlated. They identify a non-monotone joining strategy that leads to the "sputtering equilibria". Debo et al. (2012) further introduce heterogeneity in agents' private signal precision, such that a customer could be either informed or uninformed about the service quality. They show that the equilibrium joining strategy has a "hole", i.e., some customers balk at a particular queue length. In addition, by endogenizing the service rate decision, they find that a high-quality firm may choose a slow service rate to signal quality by long queue. Debo et al. (2012) explore further along this line, and study the signaling effects both by queue lengths and prices. When most customers are informed, separating equilibrium is optimal in terms of prices, and thus, "signaling by prices" dominates "signaling by queue lengths". Our fully rational model differs from the literature by incorporating an additional dimension of heterogeneity in terms of customers' delay sensitivity. Then, we further deviate from the fully rational benchmark to incorporate bounded rationalities due to the *projection effects*.

There is an emerging pool of literature on the bounded rationalities in the service operations system. Huang et al. (2013) propose queueing models with both observable and unobservable queue length, where the customers make queue-joining decisions by *probit choice*. By characterizing the quantal response equilibrium (QRE), they evaluate the impact of customers' bounded rationalities on both revenue and social welfare. Huang and Chen (2015) propose another queueing model with unobservable queue length, in which the customers learn about the waiting time by *anecdotal reasoning*, i.e., they form biased estimators of the expected waiting time by sampling from its distribution generated by the previous population. Due to limited sampling and imperfect learning about the waiting time, the customers become insensitive to the prices. Yang et al. (2014) consider a service system in which the customers are loss averse towards both price and delay attributes. Consequently, loss aversion polarizes the queues, i.e., making long queues even longer and short queues even shorter. While this chapter offers an alternative explanation towards herding, we pursue the observational learning approach, by which the polarization effect itself could be rationalized. Cui and Veeraraghavan (2014) study the "blind queues", in which customers make joining decisions based on the heterogeneous beliefs about the unknown service rate. In a sense, the type of bounded rationalities they consider are similar to ours, but we are dealing with quality uncertainty with known service rate. In addition, we identify a class of non-monotone joining strategies that are fundamentally different from the threshold strategies that they consider.

In terms of the projection effects, we learn from the literature on information biases in psychology and economics literature. The so-called "quasi-Bayesian" models refer to the mis-prediction of the state-of-the-world, which leads to inconsistency in beliefs. For example, Rabin and Schrag (1999) document how confirmation bias leads to overconfidence in belief. Madarász (2011) use the term *information projection*, which means that the agents over-estimate the probabilities that their private signals are available to others. Eyster and Rabin (2005) study an alternative setting such that the agents under-estimate the degree to which their actions are correlated with other agents' information. Eyster and Rabin (2010)

propose a model of "naive herding", where agents inadvertently over-weight early movers' private signals by neglecting that the interim herders' actions also embed these signals. Gagnon-Bartsch (2014) introduces "taste projection" in social learning, where agents over-estimate the commonness of their own taste. Empirically, this assumption is supported by social psychology studies. Ross et al. (1977), are among the early researchers who propose the "false-consensus effects" in psychology. Marks and Miller (1987) review those early studies. Reversely, the opposite direction of such egocentric bias is also documented. For example, Frable (1993) studies the "psychological marginality" effects. We would follow the definition of the *projection effects* in Gagnon-Bartsch (2014) by characterizing the first-order and the second-order bounded rationalities. In other words, we assume that agents mistakenly estimate the distribution of *types*, and further mis-predict others' estimations of the distribution. We also extend their definition, by considering both the false-consensus effects and the psychological marginality effects in the service operations systems.

5.3 Model

Service provider. We consider the market for a service of quality V_φ, $\varphi \in \{H, L\}$, and $V_H > V_L$. At the very beginning, nature flips a coin such that the service quality is V_H with probability π_0 and V_L with probability $1 - \pi_0$. The quality of the service is exogenous, and is unobserved by the customers. Regardless of the service quality, the service time is exponentially distributed with the mean $1/\mu$. The service rate μ is public information and predetermined exogenously.

Customers. Risk-neutral customers (she) arrive at the market according to a Poisson process with parameter Λ. Customers are heterogeneous both in terms of delay sensitivity (patience) and information precision about the service quality (private signals). Customers' *types* are two-dimensional attributes that include both information precision and delay sensitivity. The two attributes are independent. A proportion β of the customers are informed about the true service quality. The remaining $1 - \beta$ customers are uninformed. In other words, customers receive private signals about the service quality that are either completely informative or completely fuzzy. The proportion β is public information. The delay sensitivity is characterized by unit delay costs C_θ, $\theta \in \{H, L\}$. We assume that $C_H > C_L$, i.e., C_H refers to the impatient customers and C_L refers to the patient customers. Throughout the chapter, we use "patience" and "delay sensitivity" interchangeably. We assume that a proportion γ of all customers are impatient, i.e., H-type, while the remaining $1 - \gamma$ are patient, i.e., L-type. The true value of γ is unclear for customers, but known *ex ante* to the service provider. Customers have different estimations for γ. Such an assumption is not uncommon in the literature, e.g., Eliaz and Spiegler (2008).

Projection effects. In line with the literature, e.g., DeMarzo et al. (2003), Eyster and Rabin (2010); Eyster et al. (2014), and Gagnon-Bartsch (2014), we assume that customers suffer from (reversed-)projection psychological effects. We define such effects by characterizing customers' biased first- and second-order beliefs over the distribution of delay sensitivity

in the population. The first-order effects, i.e., *stochastically dominating perceptions*, are such that a type-θ customer believes that the fraction of the H-type customers in the system is $\hat{\gamma}_\theta^P$ under the projection effects, and $\hat{\gamma}_H^P \geq \gamma \geq \hat{\gamma}_L^P$. Reversely, a type-$\theta$ customer believes that the fraction of the H-type customers in the system is $\hat{\gamma}_\theta^R$ under the reversed-projection effects, and $\hat{\gamma}_H^R \leq \gamma \leq \hat{\gamma}_L^R$. Intuitively, such psychology induces the bounded rationalities such that the customers expect the others to be more (less) similar to themselves than reality under the (reversed-)projection effects.

Furthermore, we define $\hat{\gamma}_\theta^P(\theta')$ and $\hat{\gamma}_\theta^R(\theta')$ as a type-θ customer's belief of another type-θ' customer's estimate of the fraction of the H-type customers under the projection and the reversed-projection effects, respectively. The second-order assumption, i.e., *naivete*, is such that for any $\theta, \theta' \in \{H, L\}$, $\hat{\gamma}_\theta^P(\theta') = \hat{\gamma}_\theta^P$ and $\hat{\gamma}_\theta^R(\theta') = \hat{\gamma}_\theta^R$. In other words, the customers neglect that those of different delay sensitivity might form alternative perceptions for the distribution of patience.

Utilities. Next, we discuss the micro-structure of the customers' decision-making process. Customers arrive at the market and form a queue on a first-come first-served basis. The queue length is publicly observable. Customers are rational and calculate their expected utilities. If a customer balks, she obtains a reservation utility of zero. If she anticipates a nonnegative utility, she would join; otherwise, she would balk. For tractability, we assume that once she joins the queue, she may not renege. This is a typical assumption in the literature, e.g., Debo et al. (2012).

Consider an informed type-θ customer, $\theta \in \{H, L\}$, who enters the system knowing that the true service quality is V_H. She observes that there are already n customers in the system, including the one that is being served. The expected waiting time for her is $(n+1)/\mu$. Hence, her expected utility if she joins the queue is:

$$W_i(n, \theta, V_H) = V_H - (n+1)C_\theta/\mu, \forall \theta \in \{H, L\}. \tag{5.1}$$

Suppose an uninformed type-θ customer, $\theta \in \{H, L\}$, enters when there are already n customers in the system. Let $\hat{\alpha}_\theta^P(n)$ be her biased belief under the projection effects, i.e., her posterior estimates of the probabilities that the service quality is high. Her expected utility from joining the service is:

$$\begin{aligned} W_u(n, \theta, \hat{\alpha}_\theta^P) &= \hat{\alpha}_\theta^P(n)V_H + (1 - \hat{\alpha}_\theta^P(n))V_L - (n+1)C_\theta/\mu \\ &= \hat{\alpha}_\theta^P(n)W_i(n, \theta, V_H) + (1 - \hat{\alpha}_\theta^P(n))W_i(n, \theta, V_L), \forall \theta \in \{H, L\}. \end{aligned} \tag{5.2}$$

Similarly, we use $\alpha_\theta(n)$ or $\hat{\alpha}_\theta^R$ for the corresponding posterior beliefs when the customers are fully rational or with the reversed-projection effects, $\forall \theta \in \{H, L\}$.

In general, a customer's strategies are mappings from states to actions. Under a binary representation, we use "1" to indicate the action "join", and "0" to indicate "balk". For an informed type-θ customer, $\sigma_\theta^i : \{V_H, V_L\} \times N \to [0, 1]$. For an uninformed customer, $\sigma_\theta^u : N \to [0, 1]$. In this chapter, we mainly focus on the pure strategy equilibria, i.e., $\sigma_\theta^i, \sigma_\theta^u \in \{0, 1\}$. Table 5.1 summarizes the notations in the chapter.

Table 5.1: Summary of notations.

V_φ	The service quality, $\varphi \in \{H, L\}$, and $V_H > V_L$.
π_0	Probability of high service quality.
μ	The service rate.
Λ	The intensity of the Poisson arrival process.
C_θ	Unit delay costs, $\theta \in \{H, L\}$, and $C_H > C_L$.
β	The proportion of the informed customers.
γ	The proportion of the impatient customers.
$\hat{\gamma}_\theta^P, \hat{\gamma}_\theta^R$	The proportion of the impatient customers estimated by a type-θ customer with the (reversed-)projection bias, $\theta \in \{H, L\}$.
$\hat{\gamma}_\theta^P(\theta'), \hat{\gamma}_\theta^R(\theta')$	A type-θ customer's belief with the (reversed-)projection bias on another type-θ' customer's estimate of the proportion of the impatient customers, $\theta, \theta' \in \{H, L\}$.
$\sigma_\theta^i, \sigma_\theta^u$	Action of the informed (uninformed) type-θ customer.
$W_i(n, \theta, V_\varphi)$	The payoff of an informed type-θ customer, if she joins at queue length n, when the true quality is V_φ, $\varphi \in \{H, L\}$, $\theta \in \{H, L\}$.
$\alpha_\theta(n)$ $\hat{\alpha}_\theta^P(n), \hat{\alpha}_\theta^R(n)$	A rational (or biased, respectively) uninformed type-θ customer's posterior estimates of the probabilities that the service quality is high, when the queue length is n.
$W_u(n, \theta, \alpha(n))$	The payoff of an uninformed type-θ customer, if she joins at queue length n, with a generic posterior belief $\alpha(n)$, $\varphi \in \{H, L\}$, $\theta \in \{H, L\}$.
$\pi(n, V_\varphi, \sigma_\theta^i, \sigma_\theta^u)$ $\hat{\pi}_\theta^P, \hat{\pi}_\theta^R$	The limiting distribution for queue length in the fully rational benchmark (or biased, respectively) with service quality V_φ and given actions $\sigma_\theta^i, \sigma_\theta^u$, $\varphi \in \{H, L\}$, $\theta \in \{H, L\}$.
$WL(V_\varphi, \pi)$	Welfare loss rate given queue length distribution π, when the service quality is V_φ, $\varphi \in \{H, L\}$.

5.4 Analysis

We begin the analysis by characterizing the equilibria when the customers are fully rational. We consider equilibria in which the strategies of a customer depend on her delay sensitivity, her private signal, and the queue length. The customers form beliefs about the quality of service based on the number of customers that have preceded her, but not the complete history of her predecessors' behaviors. Then, we shall deviate from the fully rational benchmark and study the projection effects.

Fully rational benchmark

We begin our analysis for the fully rational model, where customers correctly estimate the distribution of the H-type and the L-type customers in the population. For the moment, we need the following assumptions to simplify the analysis:

Assumption 1 *Customers share common prior belief π_0 about the quality of the service ex ante.*

Assumption 2 $\left\lfloor \frac{\mu V_L}{C_H} \right\rfloor < \left\lfloor \frac{\mu V_H}{C_H} \right\rfloor < \left\lfloor \frac{\mu V_L}{C_L} \right\rfloor < \left\lfloor \frac{\mu V_H}{C_L} \right\rfloor$.

Assumption 3 $V_L \geqslant \frac{C_H}{\mu}$.

By the second assumption, we restrict our discussion to an interesting scenario such that the heterogeneity in delay sensitivity is highlighted in lieu of quality difference. Finally, the third assumption requires that the customers would join the system when the server is idle, even if the service quality is low. This guarantees that the service is always of value if we exclude the waiting cost.

For the given joining strategies, the induced queueing system follows a birth-and-death process, whose limiting distribution is given by the following lemma:

Lemma 6 *Let $\pi(n, V_\varphi, \widehat{\sigma}_\theta^i, \widehat{\sigma}_\theta^u)$ be the probability estimated by a fully rational uninformed customer that a service provider of quality V_φ sees n customers awaiting in the system, anticipating that the customers' strategy profile are $(\widehat{\sigma}_\theta^i, \widehat{\sigma}_\theta^u)$. Then:*

$$\pi(n, V_\varphi, \widehat{\sigma}_\theta^i, \widehat{\sigma}_\theta^u) = \begin{cases} \left[1 + \sum_{k=1}^{\infty} \prod_{m=0}^{m=k-1} \frac{\hat{\Lambda}(m, V_\varphi)}{\mu}\right]^{-1}, n = 0 \\ \pi(0, V_\varphi, \widehat{\sigma}_\theta^i, \widehat{\sigma}_\theta^u) \prod_{k=0}^{k=n-1} \frac{\hat{\Lambda}(k, V_\varphi)}{\mu}, n = 1, ..., \infty \end{cases}, \tag{5.3}$$

for $\forall \theta \in \{H, L\}, \forall \varphi \in \{H, L\}$, where

$$\hat{\Lambda}(n, V_\varphi) = \gamma\beta\Lambda\hat{\sigma}_H^i(n, V_\varphi) + (1-\gamma)\beta\Lambda\hat{\sigma}_L^i(n, V_\varphi) + \gamma(1-\beta)\Lambda\hat{\sigma}_H^u(n) + (1-\gamma)(1-\beta)\Lambda\hat{\sigma}_L^u(n). \tag{5.4}$$

From the birth-death process, we know that the equilibrium queue length distributions geometrically decay in the ratios of the arrival rates and the service rates. The arrival rates are obtained by summing up the four customers' segments, wherein each segment is accounted for if the corresponding action is "join". Furthermore, the anticipated strategy profile $(\widehat{\sigma}_\theta^i, \widehat{\sigma}_\theta^u)$ should be consistent with the equilibrium strategy profile $(\sigma_\theta^i, \sigma_\theta^u)$. Formally, we employ the following equilibrium concept:

Definition 1 *A Markov-perfect Bayesian equilibrium is defined by a strategy profile that satisfies the following properties:*

1. *Customers maximize their expected payoffs:*

$$\sigma_\theta^i(n, V_H) \in \arg\max_{\sigma\in[0,1]} \sigma W_i(n, \theta, V_H);$$

$$\sigma_\theta^i(n, V_L) \in \arg\max_{\sigma\in[0,1]} \sigma W_i(n, \theta, V_L), \forall\theta \in \{H, L\}; \tag{5.5}$$

$$\sigma_\theta^u(n) \in \arg\max_{\sigma\in[0,1]} \sigma W_u(n, \theta, \alpha_\theta), \forall\theta \in \{H, L\}.$$

2. *The beliefs are updated according to Bayes' rule:*

$$\alpha_H(n) = \alpha_L(n) = \frac{\pi_0\pi(n, V_H, \sigma_\theta^i, \sigma_\theta^u)}{\pi_0\pi(n, V_H, \sigma_\theta^i, \sigma_\theta^u) + (1-\pi_0)\pi(n, V_L, \sigma_\theta^i, \sigma_\theta^u)}, \tag{5.6}$$

whenever the denominator is strictly positive.

Proposition 26 *Such Markov-perfect Bayesian equilibrium exists in our game.*

In what follows, we restrict ourselves to the pure strategy equilibria to deliver intuitive results. The equilibrium pure strategies are characterized by the following proposition:

Proposition 27 *The equilibrium pure strategies of the rational customers are as follows:*

1. *The informed customers adopt a threshold strategy, i.e.,* $\sigma_\theta^i(V_\varphi, n) = 1$ *for* $n \leqslant \left\lfloor \frac{\mu V_\varphi}{C_\theta} \right\rfloor - 1$, *and* $\sigma_\theta^i(V_\varphi, n) = 0$ *for* $n > \left\lfloor \frac{\mu V_\varphi}{C_\theta} \right\rfloor - 1, \forall\theta \in \{H, L\}, \forall\varphi \in \{H, L\}$;

2. *A unique integer* $n_L^* \in \left[\left\lfloor \frac{\mu V_L}{C_L} \right\rfloor, \left\lfloor \frac{\mu V_H}{C_L} \right\rfloor - 1 \right]$ *, characterizes the equilibrium strategy of the uninformed L-type customers:*

$$\sigma_L^u(V_\varphi, n) = \begin{cases} 1, n < n_L^*, \forall\varphi \in \{H, L\} \\ 0, n = n_L^*, \forall\varphi \in \{H, L\} \\ \sigma_L^i(V_\varphi, n), n_L^* < n \leqslant \left\lfloor \frac{\mu V_H}{C_L} \right\rfloor - 1, \forall\varphi \in \{H, L\} \end{cases}; \tag{5.7}$$

3. *A set of integers* $\left\{ n_H^s \in \left[\left\lfloor \frac{\mu V_L}{C_H} \right\rfloor, \left\lfloor \frac{\mu V_H}{C_H} \right\rfloor - 1 \right], s \in S \right\}$, *where* $s = 1, 2, ..., |S|$ *characterize the equilibrium strategy of the uninformed H-type customers:*

$$\sigma_H^u(V_\varphi, n) = \begin{cases} 1, n \leqslant n_H^{|S|}, n \neq n_H^s, \forall s < |S|, \forall\varphi \in \{H, L\} \\ 0, n = n_H^s, \forall s < |S|, \forall\varphi \in \{H, L\} \\ 0, n > n_H^{|S|}, \forall\varphi \in \{H, L\} \end{cases}. \tag{5.8}$$

For the informed customers, there is no uncertainty with respect to the service quality, while the congestion costs are increasing in the queue lengths. By a straightforward cost-benefit analysis, we conclude that their equilibrium joining strategy is the "threshold strategy". Such a structure has been studied extensively in the literature, e.g., Hassin and Haviv (2003).

For an uninformed type-θ customer, if the queue length $n \leqslant \left\lfloor \frac{\mu V_L}{C_\theta} \right\rfloor - 1$, the value of service would cover the waiting cost even if the service quality is low. On the other hand, if $n > \left\lfloor \frac{\mu V_H}{C_\theta} \right\rfloor - 1$, the waiting cost would offset the service value even if the service quality is high. When $\left\lfloor \frac{\mu V_L}{C_\theta} \right\rfloor \leqslant n < \left\lfloor \frac{\mu V_H}{C_\theta} \right\rfloor$, the beliefs of those uninformed customers are crucial for their decision process. For this region, the queue length dynamics play two roles in the service systems. A long queue indicates the negative externalities among customers due to congestions, but it also signals good quality. Thus, our model captures rich interplays between incentive and information externalities.

The key insight from the equilibrium structure is how the uninformed customers learn about the service quality from the behaviors of the informed customers. To understand the intuitions, we first analyze the uninformed L-type customers. Suppose that there are n_L^* customers awaiting in the system, at which point the uninformed L-type customers would not join the queue. Then, any queue length strictly above n_L^* can only be reached if an informed customer joins the queue at n_L^*, but she would only do so if the service quality is high. Therefore, observing that the queue length is strictly above n_L^*, the uninformed L-type customers perfectly learn about the service quality. On the other hand, the informed customers would not join the service when the queue length is n_L^*, if the service quality is low. Observing that the queue length stays below n_L^*, the uninformed L-type customers infer that the service quality is low. In both cases, $\sigma_L^u(V_\varphi, n) = \sigma_L^i(V_\varphi, n)$, for $n_L^* < n \leqslant \left\lfloor \frac{\mu V_H}{C_L} \right\rfloor - 1, \forall \varphi \in \{H, L\}$. This is known as the *herding* behavior, e.g., Chamley (2004), since an uninformed customer would imitate the behaviors of the informed customers.

The surprising "hole" structure when the queue length is n_L^* is first documented by Debo et al. (2012). We would provide additional intuitions to justify the normative feature of this phenomenon. An uninformed L-type customer's expected service value would be a convex combination of V_L and V_H. This indicates that she would stop joining the queue at some threshold n_L^*. In addition, she expects that the other uninformed L-type customers would do the same. Thus, the strategy of those informed L-type customers becomes a precise signal for the service quality, which is otherwise buried in the "noise" of the other uninformed customers' behaviors.

The joining strategy of the uninformed H-type customers differs from that of the L-type in that it has multiple "holes". This is due to noisy signaling: The L-type customers would join when the queue length is between $\left\lfloor \frac{\mu V_L}{C_H} \right\rfloor$ and $\left\lfloor \frac{\mu V_H}{C_H} \right\rfloor$, regardless of the quality. Thus, each "hole" at n_H^s, $\forall s < |S|$, increases the signal precision so that the uninformed H-type customers have stronger beliefs that the quality is high, but the queue length signal

is never perfect. When the queue length is $n_H^{|S|}$, the signal is no longer strong enough to convince them to join. Intuitively, we call such non-monotone queue-joining behavior as *rational hesitation*: The uninformed and impatient customers constantly form consensus to stop joining the service, hoping to learn from the informed customers' behaviors, until they cannot be further convinced. In other words, the uninformed and impatient customers cannot completely learn about the service quality from the queue length, because the signal is obfuscated by the patient customers. Thus, our model rationalizes such complex mental struggles as equilibrium outcomes.

Projection effects

To simplify the analysis and highlight the major intuitions, we consider the extreme case such that each customer believes that others are of the same type as she is. By this extreme world view, we could simplify the first-order assumption of the *stochastically dominating perceptions*, such that $\hat{\gamma}_H^P = 1$, $\hat{\gamma}_L^P = 0$. In addition, the second-order assumption, i.e., *naivete*, requires that $\hat{\gamma}_\theta^P(\theta') = \hat{\gamma}_\theta^P$, $\forall \theta, \theta' \in \{H, L\}$. We summarize the projection effects as follows:

Assumption 4 *$\hat{\gamma}_H^P = 1$, $\hat{\gamma}_L^P = 0$; $\hat{\gamma}_\theta^P(\theta') = \hat{\gamma}_\theta^P$, $\forall \theta, \theta' \in \{H, L\}$.*

Therefore, under projection effects with naivete, the posterior beliefs about the service quality are type-dependent:

$$\begin{cases} \hat{\alpha}_H^P(n) = \frac{\pi_0 \hat{\pi}_H^P(n, V_H, \widehat{\sigma}_H^i, \widehat{\sigma}_H^u)}{\pi_0 \hat{\pi}_H^P(n, V_H, \widehat{\sigma}_H^i, \widehat{\sigma}_H^u) + (1-\pi_0)\hat{\pi}_H^P(n, V_L, \widehat{\sigma}_H^i, \widehat{\sigma}_H^u)} \\ \hat{\alpha}_L^P(n) = \frac{\pi_0 \hat{\pi}_L^P(n, V_H, \widehat{\sigma}_L^i, \widehat{\sigma}_L^u)}{\pi_0 \hat{\pi}_L^P(n, V_H, \widehat{\sigma}_L^i, \widehat{\sigma}_L^u) + (1-\pi_0)\hat{\pi}_L^P(n, V_L, \widehat{\sigma}_L^i, \widehat{\sigma}_L^u)} \end{cases}, \tag{5.9}$$

where $\hat{\pi}_\theta^P(n, V_\varphi, \sigma_\theta^i, \sigma_\theta^u)$ is the limiting probability that a service provider of quality V_φ observes n customers awaiting in the system, estimated by an uninformed type-θ customer who suffers from projection effects. Notice that it only depends on the anticipated strategy profile of customers with the same delay sensitivity. The limiting distribution is given by the birth-death process:

$$\hat{\pi}_\theta^P(n, V_\varphi, \widehat{\sigma}_\theta^i, \widehat{\sigma}_\theta^u) = \begin{cases} \left[1 + \sum_{k=1}^{\infty} \prod_{m=0}^{m=k-1} \frac{\beta\Lambda\widehat{\sigma}_\theta^i(V_\varphi, P, m) + (1-\beta)\Lambda\widehat{\sigma}_\theta^u(P,m)}{\mu}\right]^{-1}, n = 0 \\ \hat{\pi}_\theta^P(0, V_\varphi, \widehat{\sigma}_\theta^i, \widehat{\sigma}_\theta^u) \prod_{k=0}^{k=n-1} \frac{\beta\Lambda\widehat{\sigma}_\theta^i(V_\varphi, P, k) + (1-\beta)\Lambda\widehat{\sigma}_\theta^u(P,k)}{\mu}, n = 1, ..., \infty \end{cases}, \tag{5.10}$$

for $\forall \theta \in \{H, L\}, \forall \varphi \in \{H, L\}$.

The equilibrium queue length distribution is derived the same way as in (5.3). The anticipated strategy profile $(\widehat{\sigma}_\theta^i, \widehat{\sigma}_\theta^u)$ should also be consistent with the equilibrium strategy profile $(\sigma_\theta^i, \sigma_\theta^u)$. In what follows, we study the *naive quasi-Bayesian equilibrium*, which modifies the rational benchmark by incorporating the projection effects. It is *naive* because the customers

believe that everyone share the same population distribution, due to Assumption 4. It is *quasi-Bayesian* because each customer maximizes her expected payoff by putatively correct Bayesian updated belief, based on the aforementioned first- and second-order perceptions. Furthermore, we need the following assumption to specify the off-equilibrium beliefs:

Assumption 5 *If the quality estimates* $(\hat{\alpha}_H^P, \hat{\alpha}_L^P)$ *are supported on the states* $n \in Z \cap [0, \overline{n}]$, *then* $\hat{\alpha}_H^P(n) = \hat{\alpha}_L^P(n) = 1$, *for* $n > \overline{n}$.

Intuitively, this assumption on the off-equilibrium beliefs means that, if the queue length is longer than expected, then the service quality must be high.

Proposition 28 *Under Assumptions* $1-5$, *the equilibrium pure strategies under the projection effects are as follows:*

1. $\sigma_\theta^i(V_\varphi, P, n) = 1$ *for* $n \leqslant \left\lfloor \frac{\mu V_\varphi}{C_\theta} \right\rfloor - 1$, *and* $\sigma_\theta^i(V_\varphi, P, n) = 0$ *for* $n > \left\lfloor \frac{\mu V_\varphi}{C_\theta} \right\rfloor - 1, \forall \theta \in \{H, L\}, \forall \varphi \in \{H, L\}$;

2. *A unique* $n_L^P \in \left[\left\lfloor \frac{\mu V_L}{C_L} \right\rfloor, \left\lfloor \frac{\mu V_H}{C_L} \right\rfloor - 1 \right]$, *characterizes the equilibrium strategy of the uninformed customers:*

$$\sigma_L^u(V_\varphi, P, n) = \begin{cases} 1, n < n_L^P, \forall \varphi \in \{H, L\} \\ 0, n = n_L^P, \forall \varphi \in \{H, L\} \\ \sigma_L^i(V_\varphi, P, n), n_L^P < n \leqslant \left\lfloor \frac{\mu V_H}{C_L} \right\rfloor - 1, \forall \varphi \in \{H, L\} \end{cases}; \tag{5.11}$$

3. *A unique* $n_H^P \in \left[\left\lfloor \frac{\mu V_L}{C_H} \right\rfloor, \left\lfloor \frac{\mu V_H}{C_H} \right\rfloor - 1 \right]$, *characterizes the equilibrium strategy of the uninformed customers:*

$$\sigma_H^u(V_\varphi, P, n) = \begin{cases} 0, n = n_H^P, \forall \varphi \in \{H, L\} \\ 1, n \leqslant \left\lfloor \frac{\mu V_H}{C_H} \right\rfloor - 1, n \neq n_H^P, \forall \varphi \in \{H, L\} \end{cases}. \tag{5.12}$$

In equilibrium, the uninformed L-type customers would join the queue until $n = \left\lfloor \frac{\mu V_H}{C_L} \right\rfloor - 1$ if the service quality is high, since the informed customers would cross the "hole" at $n = n_L^P$ for them. The uninformed L-type customers balk at $n = n_L^P$ if the service quality is low. The projection effects have little impact on the equilibrium strategy of the L-type customers, who would wait at least until $n = \left\lfloor \frac{\mu V_L}{C_L} \right\rfloor - 1$, regardless of the service quality. The assumption that $\left\lfloor \frac{\mu V_H}{C_H} \right\rfloor < \left\lfloor \frac{\mu V_L}{C_L} \right\rfloor$ would de-noise the signal of queue length by screening out the H-type customer beyond $n = \left\lfloor \frac{\mu V_H}{C_H} \right\rfloor$.

On the other hand, the uninformed H-type customers would continue to join the queue beyond the "hole" at $n = n_H^P$, until $n = \left\lfloor \frac{\mu V_H}{C_H} \right\rfloor - 1$, since the L-type customers would cross

the "hole" at $n = n_H^P$ for them regardless of the true quality. This bounded rationality is due to the putatively Bayesian belief updating using the wrong type distribution under the projection effects. The uninformed H-type customers should not join the queue longer than n_H^P when the service quality is low. However, the L-type customers mislead them into the wrong belief that the quality is high and cross the "hole" for them. Thus, the uninformed H-type customers over-estimate the signal precision by queue length under the projection effects. This is somewhat surprising, since the customers who feel painful to wait are induced to wait more often, and thus could be potentially detrimental in terms of social welfare.

Reversed-projection effects

Similar to the analysis for the projection effects, we make the following simplifying assumption, which restricts our discussion to the extreme case while incorporating *naivete*:

Assumption 6 $\hat{\gamma}_H^R = 0$, $\hat{\gamma}_L^R = 1$; $\hat{\gamma}_\theta^R(\theta') = \hat{\gamma}_\theta^R$, $\forall \theta, \theta' \in \{H, L\}$.

The uninformed type-θ customers under the reversed-projection arrive at the market, observing the queue length, update their beliefs based on the anticipated strategy profile $(\widehat{\sigma}_{\theta'}^i, \widehat{\sigma}_{\theta'}^u)$, for $\forall \theta' \neq \theta$:

$$\begin{cases} \hat{\alpha}_H^R(n) = \frac{\pi_0 \hat{\pi}_H^R(n, V_H, \widehat{\sigma}_L^i, \widehat{\sigma}_L^u)}{\pi_0 \hat{\pi}_H^R(n, V_H, \widehat{\sigma}_L^i, \widehat{\sigma}_L^u) + (1-\pi_0)\hat{\pi}_H^R(n, V_L, \widehat{\sigma}_L^i, \widehat{\sigma}_L^u)} \\ \hat{\alpha}_L^R(n) = \frac{\pi_0 \hat{\pi}_L^R(n, V_H, \widehat{\sigma}_H^i, \widehat{\sigma}_H^u)}{\pi_0 \hat{\pi}_L^R(n, V_H, \widehat{\sigma}_H^i, \widehat{\sigma}_H^u) + (1-\pi_0)\hat{\pi}_L^R(n, V_L, \widehat{\sigma}_H^i, \widehat{\sigma}_H^u)} \end{cases}, \tag{5.13}$$

where $\hat{\pi}_\theta^R(n, V_\varphi, \sigma_{\theta'}^i, \sigma_{\theta'}^u)$ is the probability estimated by an uninformed type-θ customer who suffers from the reversed-projection effects, that a service provider with quality V_φ observes n customers waiting in the system:

$$\hat{\pi}_\theta^R(n, V_\varphi, \widehat{\sigma}_{\theta'}^i, \widehat{\sigma}_{\theta'}^u) = \begin{cases} \left[1 + \sum_{k=1}^{\infty} \prod_{m=0}^{m=k-1} \frac{\beta\Lambda\widehat{\sigma}_{\theta'}^i(V_\varphi, R, m) + (1-\beta)\Lambda\widehat{\sigma}_{\theta'}^u(R,m)}{\mu}\right]^{-1}, n = 0 \\ \hat{\pi}_\theta^R(0, V_\varphi, \widehat{\sigma}_{\theta'}^i, \widehat{\sigma}_{\theta'}^u) \prod_{k=0}^{k=n-1} \frac{\beta\Lambda\widehat{\sigma}_{\theta'}^i(V_\varphi, R, k) + (1-\beta)\Lambda\widehat{\sigma}_{\theta'}^u(R,k)}{\mu}, n = 1, ..., \infty \end{cases}, \tag{5.14}$$

for $\forall \theta \in \{H, L\}, \theta' \neq \theta, \forall \varphi \in \{H, L\}$. Similar to the projection case, we also need the following assumption to specify the off-equilibrium beliefs:

Assumption 7 *If the quality estimates* $(\hat{\alpha}_H^R, \hat{\alpha}_L^R)$ *are supported on the states* $n \in Z \cap [0, \overline{n}]$, *then* $\hat{\alpha}_H^R(n) = \hat{\alpha}_L^R(n) = 1$, *for* $n > \overline{n}$.

By enforcing the anticipated strategy profile $(\widehat{\sigma}_\theta^i, \widehat{\sigma}_\theta^u)$ to be consistent with the equilibrium strategy profile $(\sigma_\theta^i, \sigma_\theta^u)$, we have the following equilibrium characterization:

Proposition 29 *Under Assumptions* $1 - 3$, 6, *and* 7, *the equilibrium strategy profile of customers under the reversed-projection effects is as follows:*

1. $\sigma_\theta^i(V_\varphi, R, n) = 1$ *for* $n \leqslant \left\lfloor \frac{\mu V_\varphi}{C_\theta} \right\rfloor - 1$, *and* $\sigma_\theta^i(V_\varphi, R, n) = 0$ *for* $n > \left\lfloor \frac{\mu V_\varphi}{C_\theta} \right\rfloor - 1, \forall \theta \in \{H, L\}, \forall \varphi \in \{H, L\}$;

2. $\sigma_L^u(V_\varphi, R, n) = 1$ *for* $n \leqslant \left\lfloor \frac{\mu V_H}{C_L} \right\rfloor - 1$;

3. *There exists threshold* $n_H^R \in \left[\left\lfloor \frac{\mu V_L}{C_H} \right\rfloor - 1, \left\lfloor \frac{\mu V_H}{C_H} \right\rfloor - 1 \right]$, *such that* $\sigma_H^u(V_\varphi, R, n) = 1$ *for* $n \leqslant n_H^R$ *and* $\sigma_H^u(V_\varphi, R, n) = 0$ *for* $n > n_H^R$.

The informed customers adopt the same queue-joining strategy as in the fully rational benchmark. The uninformed $L-$type customers believe that all customers are of H-type, and would never join the queue when $n \geqslant \left\lfloor \frac{\mu V_H}{C_H} \right\rfloor$. However, the informed $L-$type customers would join when $n \leqslant \left\lfloor \frac{\mu V_L}{C_L} \right\rfloor - 1$ regardless of the service quality. Thus, the assumption on the off-equilibrium beliefs imply that the uninformed $L-$type customers would expect that the service is of high quality with probability one and they would join when $n \leqslant \left\lfloor \frac{\mu V_H}{C_L} \right\rfloor - 1$. In other words, the uninformed $L-$type customers over-estimate the information precision. Intuitively, the patient customers tend to under-estimate other customers' patience, and thus over-estimate the service quality. Consequently, too many patient customers would join the queue when they should not.

While the above discussions reveal that the reversed-projection effects seem to select the right customers to wait in the queue, they give rise to a different source of inefficiency: For the uninformed $H-$type customers, a threshold strategy is optimal. Intuitively, the impatient customers over-estimate others' patience, but they learn nothing from patient customers' behaviors in their observation window. Thus, the reversed-projection effects impede social learning for the impatient customers. Note that the converse is not true under the projection effects. This asymmetry arises because of the inherent difference of joining behaviors between the two segments.

Welfare Implications

The customers' incorrect beliefs about the population distribution not only are sustained in the learning equilibria, but also have impacts on the social welfare. To summarize the welfare implications of the equilibrium queue-joining strategies, we proceed to define the equilibrium *welfare loss rates* or *regret rates*, i.e., the potential payoffs that the uninformed customers could capture if they know the service quality. We calculate the welfare loss rates by considering the *type-I decision errors*, i.e., balking from the queue when the service quality is high, and *type-II errors*, i.e., joining the queue when the service quality is low.

The welfare loss rate in the fully rational benchmark is accounted for as follows:

$$\begin{aligned} WL &= \pi_0 \left[WL(V_H, \pi(n, V_H))\right] + (1-\pi_0)\left[WL(V_L, \pi(n, V_H))\right] \\ &= \underbrace{\pi_0(1-\beta)\Lambda \cdot \sum_{n=\left\lfloor \frac{\mu V_L}{C_H} \right\rfloor}^{n=\left\lfloor \frac{\mu V_H}{C_L} \right\rfloor - 1} \left\{ \begin{array}{c} \gamma \left[V_H - (n+1)\frac{C_H}{\mu}\right] \left[\sigma_H^i(n, V_H) - \sigma_H^u(n)\right] \pi(n, V_H) \\ +(1-\gamma)\left[V_H - (n+1)\frac{C_L}{\mu}\right] \left[\sigma_L^i(n, V_H) - \sigma_L^u(n)\right] \pi(n, V_H) \end{array} \right\}}_{\text{Type-I errors when the quality is high}} \quad (5.15) \\ &\quad \underbrace{+(1-\pi_0)(1-\beta)\Lambda \cdot \sum_{n=\left\lfloor \frac{\mu V_L}{C_H} \right\rfloor}^{n=\left\lfloor \frac{\mu V_H}{C_L} \right\rfloor - 1} \left\{ \begin{array}{c} \gamma \left[(n+1)\frac{C_H}{\mu} - V_L\right] \left[\sigma_H^u(n) - \sigma_H^i(n, V_L)\right] \pi(n, V_L) \\ +(1-\gamma)\left[(n+1)\frac{C_L}{\mu} - V_L\right] \left[\sigma_L^u(n) - \sigma_L^i(n, V_L)\right] \pi(n, V_L) \end{array} \right\}}_{\text{Type-II errors when the quality is low}}. \end{aligned}$$

When the quality is high, a random uninformed customer arrives with Poisson rate $(1-\beta)\Lambda$. By the "Poisson arrival see time average" properties (Wolff, 1982), we calculate the expected loss in payoff by conditioning on the queue lengths. Thus, it suffices to discuss the two types of decision errors induced by the equilibrium queue-joining strategies with respect to the queue lengths. Similarly, we can perform the same accounting under the (reversed-)projection effects.

Table 5.2 summarizes the two types of decision errors with respect to the queue lengths that an arriving customer observes. In particular, when the queue length $n \in \left[n_H^{|S|} + 1, \left\lfloor \frac{\mu V_H}{C_H} \right\rfloor\right)$, the uninformed and impatient customers always join the queue under the projection effects. Thus, compared with the fully rational benchmark, they reduce the welfare loss due to the type-I errors when the service quality is high, while increasing the welfare loss due to the type-II errors when the service quality is low. Similarly, when the queue length $n \in \left[\left\lfloor \frac{\mu V_L}{C_L} \right\rfloor, \left\lfloor \frac{\mu V_H}{C_L} \right\rfloor\right)$, the uninformed and patient customers always join the queue under the reversed-projection effects. Thus, the inefficiency of social learning driven by such bounded rationalities could potentially impair social welfare.

5.5 Numerical Examples

In this section, we provide some numerical examples to support our intuitions. Consider a service system where $V_H = 18$, $V_L = 10$, $\mu = 1$, $C_H = 0.4$, $C_L = 0.18$, $\gamma = 0.8$, $\beta = 0.2$, $\pi_0 = 0.18$, and $\lambda = 0.6$. For the moment, we assume that the customers are fully rational. Figure 5.1 plots the likelihood ratio function with respect to the queue length. Since a smaller likelihood ratio indicates a higher probability of good quality, this figure illustrates that "long queue signals high quality".

Figure 5.2 plots the benefits and the costs for joining the service system. The blue curve

Table 5.2: The decision errors with respect to queue lengths.

customer types	uninformed and impatient		uninformed and patient	
service quality	high quality	low quality	high quality	low quality
error types	type-I errors	type-II errors	type-I errors	type-II errors
fully rational	$n = n_H^s, \forall s$,and $\left[n_H^{\lvert S\rvert}+1, \left\lfloor \frac{\mu V_H}{C_H} \right\rfloor\right)$	$\forall n \in \left[\left\lfloor \frac{\mu V_L}{C_H} \right\rfloor, n_H^{\lvert S\rvert}\right)$, and $n \neq n_H^s, \forall s$	n_L^*	$\forall n \in \left[\left\lfloor \frac{\mu V_L}{C_L} \right\rfloor, \left\lfloor \frac{\mu V_H}{C_L} \right\rfloor\right)$, and $n \neq n_L^*$
projection effects	n_H^P	$\forall n \in \left[\left\lfloor \frac{\mu V_L}{C_H} \right\rfloor, \left\lfloor \frac{\mu V_H}{C_H} \right\rfloor\right)$, and $n \neq n_H^P$	n_L^P	$\forall n \in \left[\left\lfloor \frac{\mu V_L}{C_L} \right\rfloor, \left\lfloor \frac{\mu V_H}{C_L} \right\rfloor\right)$, and $n \neq n_L^P$
reversed-projection	$\forall n \in \left[n_H^R+1, \left\lfloor \frac{\mu V_H}{C_H} \right\rfloor\right)$	$\forall n \in \left[\left\lfloor \frac{\mu V_L}{C_H} \right\rfloor, n_H^R\right]$	$\varnothing$	$\forall n \in \left[\left\lfloor \frac{\mu V_L}{C_L} \right\rfloor, \left\lfloor \frac{\mu V_H}{C_L} \right\rfloor\right)$

Note: In the last three rows, each entry shows the queue lengths at which an entering customer makes decision error of certain type, depending on her attributes and the service quality.

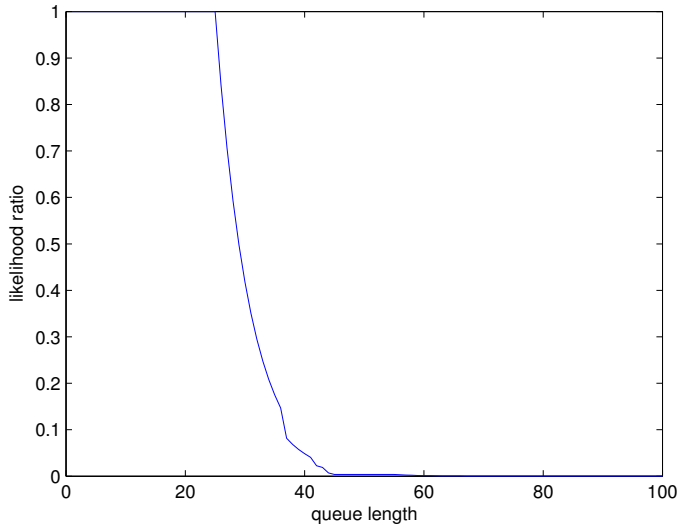

Figure 5.1: Likelihood ratio decreases in queue length.

is the *virtual valuation*, which is defined as:

$$\hat{V}(n) = \frac{\pi_0}{\pi_0 + (1-\pi_0) l(n, \sigma_\theta^i, \sigma_\theta^u)} V_H + \frac{(1-\pi_0)}{\pi_0 / l(n, \sigma_\theta^i, \sigma_\theta^u) + (1-\pi_0)} V_L, \tag{5.16}$$

which intuitively represents the value of service in expectation. We draw its continuous interpolation instead of the piece-wise integer stairs. From the figure we can see that the underlying virtual valuation function is non-convex, non-concave, and it contains plateaus.

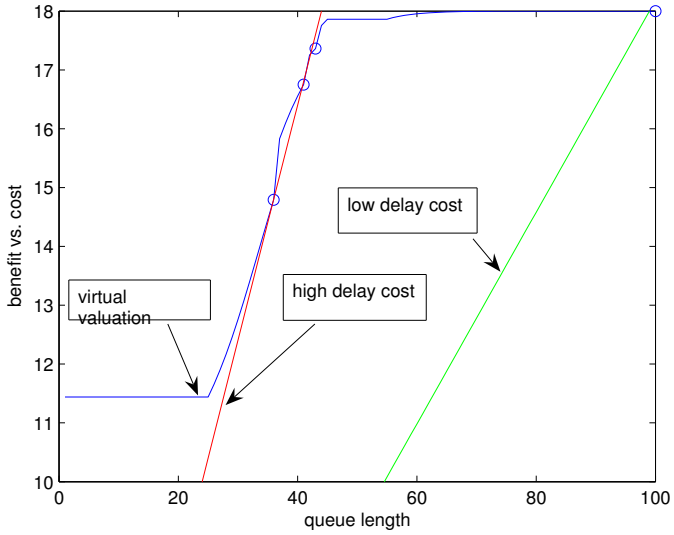

Figure 5.2: The benefit and the cost of joining the service.

The red line and the green line represent the delay costs for the impatient and the patient customers, respectively. Whenever the virtual valuation falls below the delay costs, we would expect a "hole" or a threshold in the strategy space. Furthermore, since the queue length is long in this example, the patient customers receive a clear signal of the quality and join the system until $n = \left\lfloor \frac{\mu V_H}{C_H} \right\rfloor$.

Table 5.3: The positions of "holes" change with the fraction of the impatient customers (γ).

γ	n_H^1	n_H^2	n_H^3	n_H^4
0.72	32	37	40	43
0.76	34	40	42	—
0.80	36	41	43	—
0.84	39	42	—	—
0.88	41	43	—	—

From Table 5.3 we can see that, the positions of the "holes" would be shifted backward with the increasing fractions of the impatient customers. From Table 5.4 we can see that, the positions of the "holes" would be shifted backward with the increasing fractions of the informed customers. Intuitively, the uninformed H-type customers form belief based on the difference of the equilibrium behaviors of informed H-type customers facing high quality and low quality service, while the L-type customers are the noisy part of the queue length signal. Thus, with the increasing fraction of H-type customers, the queue length signals quality more clearly. Similarly, with the increasing fraction of the informed customers, the signal precision also improves. Now that the virtual valuation would be higher when an entering uninformed customer observes the queue, she is more likely to join. Therefore, the positions of "holes" would be shifted backwards.

Table 5.4: The positions of "holes" change with the fraction of the informed customers (β).

β	n_H^1	n_H^2	n_H^3	n_H^4
0.17	31	37	41	44
0.18	32	39	42	—
0.19	34	40	42	—
0.20	36	41	43	—
0.21	39	42	—	—

Consider another service system where $V_H = 18$, $V_L = 10$, $\mu = 1$, $C_H = 0.4$, $C_L = 0.18$, $\gamma = 0.84$, $\beta = 0.1$, $\pi_0 = 0.33$, and $\lambda = 0.99$. In Figure 5.3 we compare the cumulative distributions for queue lengths when the H-type customers are fully rational, under projection effects, and under reversed-projection effects. Note that the L-type customers' belief does not matter in this example, as all the uninformed L-type customers would join as if they know that the service quality is high. For given psychological effects, the queue lengths when the true quality is high stochastically dominates that when the true quality is low. For given underlying service quality, the queue lengths are the longest when the H-type customers suffer from projection effects, and shortest when they are fully rational.

In Table 5.5 we compare the positions of "holes" or thresholds as well as the expected queue lengths under different psychological effects. The queue length results are consistent with the cumulative distribution curves. Longer queue lengths indicate higher system utilization, and thus lead to higher rate of value creation.

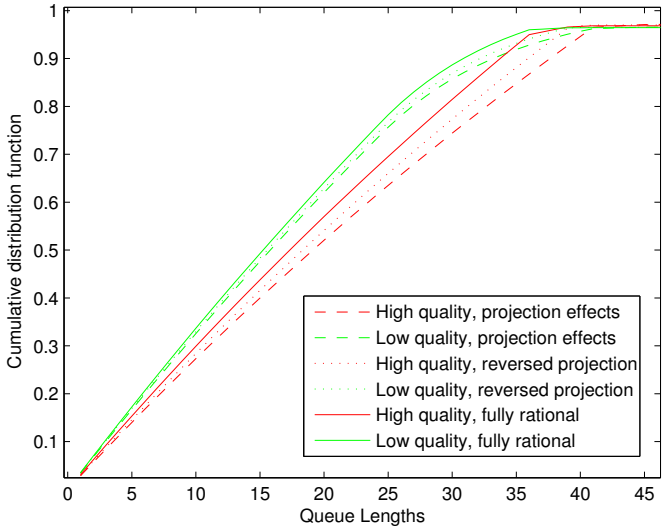

Figure 5.3: Comparison of cumulative distributions for queue lengths.

Table 5.5: Comparison of system performances under different psychological effects.

Psychological effects	Fully rational	Projection effects	Reversed projection
Position of holes/thresholds	36, 39, 41	41	39
Expected queue lengths (high quality)	17.25	19.17	18.31
Expected queue lengths (low quality)	15.27	16.07	15.72

5.6 Mixed models and performances comparison

We have characterized the equilibria when all the customers suffer from either the projection or the reversed-projection effects. To understand the impacts of the projection effects on the system performances, we summarize the characterizations of the other intermediate cases for completeness.

Proposition 30 *Suppose that the H-type customers suffer from the projection effects, while the L-type customers suffer from the reversed-projection effects, i.e., $\hat{\gamma}_H^P = \hat{\gamma}_L^R = 1$. Equipped with the Assumptions 1-3, 5, 7, and* naivete*, the equilibrium queue-joining strategies are as follows.*

- *For the informed customers:*

  1. $\sigma_H^i(V_\varphi, P, n) = 1$ *for* $n \leqslant \left\lfloor \frac{\mu V_\varphi}{C_H} \right\rfloor - 1$, *and* $\sigma_H^i(V_\varphi, P, n) = 0$ *for* $n > \left\lfloor \frac{\mu V_\varphi}{C_H} \right\rfloor - 1, \forall \varphi \in \{H, L\}$;

2. $\sigma_L^i(V_\varphi, R, n) = 1$ *for* $n \leqslant \left\lfloor \frac{\mu V_\varphi}{C_L} \right\rfloor - 1$, *and* $\sigma_L^i(V_\varphi, R, n) = 0$ *for* $n > \left\lfloor \frac{\mu V_\varphi}{C_L} \right\rfloor - 1, \forall \varphi \in \{H, L\}$.

- *For the uninformed customers:*

1. *A unique* $n_H^P \in \left[\left\lfloor \frac{\mu V_L}{C_H} \right\rfloor, \left\lfloor \frac{\mu V_H}{C_H} \right\rfloor - 1 \right]$, *characterizes the equilibrium behaviors of the uninformed H-type customers:*

$$\sigma_H^u(P, n) = \begin{cases} 1, n \leqslant \left\lfloor \frac{\mu V_H}{C_H} \right\rfloor - 1, n \neq n_H^P \\ 0, n = n_H^P \end{cases}; \tag{5.17}$$

2. $\sigma_L^u(R, n) = 1$ *for* $n \leqslant \left\lfloor \frac{\mu V_H}{C_L} \right\rfloor - 1$, *and* $\sigma_L^u(R, n) = 0$ *for* $n > \left\lfloor \frac{\mu V_H}{C_L} \right\rfloor - 1$.

The results suggest that both the impatient and the patient customers simultaneously underestimate others' patience, and join the long queue when the service quality is low. One the other hand, we summarize the results for the alternative intermediate case.

Proposition 31 *Suppose that the H-type customers suffer from the reversed-projection effects and the L-type customers suffer from the projection effects, i.e.,* $\hat{\gamma}_H^R = \hat{\gamma}_L^P = 0$. *Equipped with the Assumptions 1-3, 5, 7, and* naivete, *the equilibrium queue-joining strategies are as follows.*

- *For the informed customers:*

1. $\sigma_H^i(V_\varphi, P, n) = 1$ *for* $n \leqslant \left\lfloor \frac{\mu V_\varphi}{C_H} \right\rfloor - 1$, *and* $\sigma_H^i(V_\varphi, P, n) = 0$ *for* $n > \left\lfloor \frac{\mu V_\varphi}{C_H} \right\rfloor - 1, \forall \varphi \in \{H, L\}$;
2. $\sigma_L^i(V_\varphi, R, n) = 1$ *for* $n \leqslant \left\lfloor \frac{\mu V_\varphi}{C_L} \right\rfloor - 1$, *and* $\sigma_L^i(V_\varphi, R, n) = 0$ *for* $n > \left\lfloor \frac{\mu V_\varphi}{C_L} \right\rfloor - 1, \forall \varphi \in \{H, L\}$.

- *For the uninformed customers:*

1. *There exists a threshold* $n_H^R \in \left[\left\lfloor \frac{\mu V_L}{C_H} \right\rfloor - 1, \left\lfloor \frac{\mu V_H}{C_H} \right\rfloor - 1 \right]$, *such that* $\sigma_H^u(R, n) = 1$ *for* $n \leqslant n_H^R$ *and* $\sigma_H^u(R, n) = 0$ *for* $n > n_H^R$;
2. *A unique* $n_L^P \in \left[\left\lfloor \frac{\mu V_L}{C_L} \right\rfloor, \left\lfloor \frac{\mu V_H}{C_L} \right\rfloor - 1 \right]$, *characterizes the equilibrium behaviors of the uninformed L-type customers:*

$$\sigma_L^u(V_\varphi, P, n) = \begin{cases} 1, n < n_L^P \\ 0, n = n_L^P \\ 1, n_L^P < n \leqslant \left\lfloor \frac{\mu V_H}{C_L} \right\rfloor - 1, \varphi = H \\ 0, n_L^P < n \leqslant \left\lfloor \frac{\mu V_H}{C_L} \right\rfloor - 1, \varphi = L \end{cases}. \tag{5.18}$$

In this case, both the patient and the impatient customers over-estimate the others' patience. The queueing dynamics become uninformative for the impatient customers, and they adopt the threshold strategies, while the patient customers are less sensitive to the projection bias. In what follows, we would first define the performance measures and then compare the system performances for different models.

Definition 2 *Likelihood ratio ordering. Let $\mathcal{X}$, $\mathcal{Y}$ be two discrete random variables on a common support set $\mathcal{N}$. Then, $\mathcal{X}$ dominates $\mathcal{Y}$ in likelihood ratio ordering, denoted by $\mathcal{X} \succeq_{lr} \mathcal{Y}$, if and only if*

$$P(\mathcal{X} = n)P(\mathcal{Y} = n-1) \geq P(\mathcal{X} = n-1)P(\mathcal{Y} = n), \forall n \in \mathcal{N}. \tag{5.19}$$

Definition 3 *Stochastic dominance (first-order). Let $\mathcal{X}$, $\mathcal{Y}$ be two discrete random variables on a common support set $\mathcal{N}$. Then, $\mathcal{X}$ dominates $\mathcal{Y}$ in terms of the first-order stochastic dominance, denoted by $\mathcal{X} \succeq_{st} \mathcal{Y}$, if and only if $P(\mathcal{X} \geq n) \geq P(\mathcal{Y} \geq n), \forall n \in \mathcal{N}$.*

For notational convenience, let $\mathcal{Q}(HP, LP, V_\varphi)$ be the queue length when both the H-type and the L-type customers suffer from the projection effects when the true service quality is V_φ. Similarly, we could define $\mathcal{Q}(HP, LR, V_\varphi)$, $\mathcal{Q}(HR, LP, V_\varphi)$, and $\mathcal{Q}(HR, LR, V_\varphi)$. The following proposition compares the system performances among all four cases.

Proposition 32 *For $\forall \varphi \in \{H, L\}$, $\mathcal{Q}(HP, LR, V_\varphi) \succeq_{lr} \mathcal{Q}(HP, LP, V_\varphi)$. Furthermore, if $n_H^P \geq \left\lfloor \frac{\mu V_L}{C_H} \right\rfloor + 1 + \frac{\log(C_0)}{\log(1-\beta)}$, for some constant C_0 (whose specific expression is given in the appendix), $\mathcal{Q}(HP, LR, V_\varphi) \succeq_{lr} \mathcal{Q}(HR, LR, V_\varphi) \succeq_{lr} \mathcal{Q}(HR, LP, V_\varphi)$. The same order holds in terms of the first-order stochastic dominance and the expected performance measure $\mathrm{E}h(\cdot)$, for any nondecreasing measurement function $h(\cdot)$.*

The major result suggests that, the queue lengths are the longest when the impatient customers suffer from the projection effects, while the patient customers suffer from the reversed-projection effects. The uninformed H-type customers under the projection effects over-estimate the signal precision by queue length, and join the queue when $n > n_H^P$ due to the optimism bias. On the other hand, the patient customers under the reversed-projection effects under-estimate the signal precision by queue length, and the off-equilibrium beliefs drive them to join the long queue. In both cases, "buying frenzy" happens in service systems when customers under-estimate others' patience. As a managerial recommendation, the service provider could launch marketing campaign to convince the potential customers that other people are impatient. By forming rational beliefs that the impatient people would not wait in a congested environment, customers receive an optimistic signal about the service quality from the long queue.

5.7 Conclusion

In this chapter, we study the customers' learning behaviors in the service operations systems, when customers hold incorrect beliefs about the population distribution. By proposing a single-server queueing model with observable queue length, in which customers are heterogeneous both in terms of their delay sensitivity and private signals. In the fully rational benchmark, we identify the *rational hesitation* in the pure equilibrium queue-joining strategies, i.e., a non-monotone structure with multiple "holes". Intuitively, this means that the uninformed and impatient customers constantly form consensus to stop joining the queue, hoping to learn better from the informed customers' behaviors, which is obfuscated by the uninformed and patient customers.

Furthermore, we deviate from the fully rational benchmark to consider the cases when the customers suffer from either the *false consensus* or the *psychological marginality effects* in terms of their patience, i.e., the *projection effects* and the *reversed-projection effects*. Somewhat surprisingly, under projection bias, customers who are more averse to waiting will react more sensitively to the observed long queue, which leads to over-estimation of the service quality and waiting on the long queue. Conversely, under reversed-projection bias, the patient customers tend to under-estimate other customers' patience, over-estimate the service quality, and wait in the long queue.

Chapter 6

Conclusions

This dissertation consists of four essays on service systems, with considerations of incentives, information asymmetries and bounded rationalities. They are motivated by information service operations in agriculture, distribution of product/technology in developing economies, fresh-product delivery service, and tourism industries, respectively.

In Chapter 2, we study the incentives for farmers' cooperatives in developing economies to conglomerate and form farmer producer organizations (FPOs). We propose a stylized Cournot competition model under incomplete information and study the incentives of FPOs' formation in developing economies. We focus on the functionality of FPOs as information sharing coalition. We further distill our results by calibrating the interactions of four effects:

- *Competition effect*. Under this effect, the over-precision of private signals is detrimental towards farmers' revenues, and the public information has an adverse impact concerning farmers' aggregate payoff.
- *Congestion effect*. The value of a private signal diminishes in the number of farmers who respond to it. This effect prevents the formation of giant coalitions.
- *Crowding-out effect*. When the farmers are coordinated in terms of information acquisition, the public information substitutes the private information.
- *Polarization effect*. High public information provision leads to the *dominant group architecture*. However, the polarization in terms of private information achieves a fair allocation of social welfare among the farmers.

From the policy perspective, the identifications of those effects offer rich insights concerning the NGO's dual roles in providing market information as well as mobilizing farmers to build FPOs in the developing economies. We find this research area to be exciting because there are many unexplored issues: (1) Value of short-term market information towards better selling decisions. (2) Heterogeneity or alternative costs structure in terms of the information acquisition. (3) Other facets of FPOs', e.g., increasing bargaining power of farmers against

retailers, sharing technology information, and etc. (4) Other important factors in agriculture, e.g., yield uncertainty, multiple markets and resources constraints. We hope our research will motivate future research in this emerging area.

In Chapter 3, we propose a stylized monopoly pricing model with investment goods, wherein consumers suffer from *present-bias*: Consumers procrastinate purchase decisions but make no purchase later due to lack of self-control. We show that advance selling can be beneficial both to the seller as an inter-temporal discrimination instrument, and to the consumers as a commitment device.

We highlight some policy recommendations towards solving the product adoption puzzle: (1) Micro-finance instruments such as low loan rates might have adverse effects on product adoption. (2) Timely subsidy in the advance-market can be most efficient expenditure of the donor's funding. (3) Increasing public awareness of lack of self-control (financial responsibility education) may or may not help, as it can either increase or decrease donor's subsidy level.

In Chapter 4, we propose a model of service operations systems in which customers are heterogeneous both in terms of their private delay sensitivity and taste preference. The service provider maximizes revenue through jointly optimal pricing strategies, steady-state scheduling rules, and probabilistic routing policies under information asymmetry. The impact of *horizontal substitutions* is twofold: It provides instrument to balance the traffic intensities between horizontal differentiated services, however, the service provider should sacrifice information rent to create incentives for customers to truthfully report their taste preference.

This chapter contributes to the literature by extending the standard feasible region approach to novel model settings. In particular, the analytical results concerning the flexible customers' equilibrium queue-joining choices inspire a *hierarchical load-balancing* heuristic algorithm. The heuristic algorithm is used to solve the second-stage problem for the server-specific model, but it could be applied in the (more general) basic model for the service provider to route its customers. On the managerial side, this chapter sheds interesting light on the impact of customers' taste indifference on the service systems. Intuitively, when the demands for horizontally differentiated services are unbalanced, the customers with taste indifference are valuable since they could be used for load-balancing. However, when the demands are relatively balanced, the customers with taste indifference are less valuable since the service provider should sacrifice information rent to create incentive for them to truthfully report their taste. Consequently, if the flexible customers are treated differently from the dedicated ones, they should wait longer when the arrival rates to both queues are relatively balanced, while shorter expected delays should be assigned to them when the traffic in the two queues are unbalanced. The results thus provide rich insights on the role of information asymmetry in the interaction between the flexible customers and the dedicated ones.

In Chapter 5, we study the customers' learning behaviors in the service operations systems, when customers hold incorrect beliefs about the population distribution. We propose a single-server queueing model with observable queue length, in which the customers are heterogeneous both in terms of their delay sensitivity and information precision about the

unknown service quality. The bounded rationalities impede effective learning by inducing decision errors, which could reduce the social welfare due to "long wait for bad service". Examples of such blind "buying frenzy" are not uncommon, even if the service quality is low.

From the perspective of a service provider who maximizes the system utilization rate, the insight from our analysis leads to the managerial recommendation to exploit and manipulate the customers' incorrect beliefs about population distribution via marketing campaign. For instance, a dentist on the ZocDoc.com should emphasize on the urgent conditions of her patients, so that the public perception on her popularity could be exaggerated. In contrast, public information disclosure to disenchant the customers of their incorrect beliefs could be beneficial from the welfare perspective.

Appendix A

Appendix for Chapter 2

In this appendix, we provide the detailed proofs of the main results in Chapter 2. To streamline our analysis, we begin by proving the general results in Lemma 2.

Proof of Lemma 2. Firstly, we define a partitioning of the set $N = \{1, 2, ..., n\}$ as a m-tuple $(N_1, N_2, ..., N_m)$, such that $N_i \cap N_j = \emptyset$, and $\cup_{i=1}^{i=m} N_i = N$. Denote the cardinality of $|N_i| = n_i$. Any coalition configuration consisting of n farmers can be represented by a partitioning $(N_1, N_2, ..., N_m)$ such that any two farmers in the set N_i are connected. By restriction to linear Bayesian response, we assume that the production quantity $q_i = A_i + B_i y_i + C_i x_0$, where the signal $y_i = \frac{\sum_{j \in N_i} \gamma_j \cdot x_j}{\sum_{j \in N_i} \gamma_j}$ is the combination of signals within the set N_i. Due to additivity of Gaussian signals, we know that the precision of y_i is $\rho_i = \sum_{j \in n_i} \gamma_j$. Farmer i's revenue is

$$R_i\left(N_1, N_2, ..., N_m\right) = \left(a - b\sum_{j=1}^{m} n_j \mathbb{E}[q_j | x_0, y_i] + \mathbb{E}[u | x_0, y_i]\right) \cdot q_i - c q_i. \tag{A.1}$$

The first-order condition gives rise to:

$$\begin{aligned} A_i &= \frac{a-c}{(1+\sum_{i=1}^{m} n_i)b}, \\ B_i &= \frac{\rho_i}{\left[1+\sum_{i=1}^{m}\left(\frac{n_i \rho_i}{\alpha+\beta+\rho_i}\right)\right](\alpha+\beta+\rho_i)b}, \\ C_i &= \frac{\beta}{\left[1+\sum_{i=1}^{m}\left(\frac{n_i \rho_i}{\alpha+\beta+\rho_i}\right)\right] b} \cdot \left[\frac{1}{\alpha+\beta+\rho_i} - \frac{\sum_{i=1}^{m} \frac{n_i}{\alpha+\beta+\rho_i}}{1+\sum_{i=1}^{m} n_i}\right]. \end{aligned} \tag{A.2}$$

Thus, $R_i\left(N_1, N_2, ..., N_m\right) = \mathbb{E}\left\{\mathbb{E}\left[q_i^*\left(N_1, N_2, ..., N_m\right)^2 b | x_0, y_i\right]\right\}$, and the result follows. □

Proof of Proposition 1. Firstly, we summarize the possible equilibria in Table A.1 following the standard calculation procedure. When $g_{12} = g_{21} = 1$, both farmers have incentive to disconnect to save the connection cost. The only other possible equilibrium is when $\gamma_1^* = \alpha - \beta$, $\gamma_2^* = 0$, and $g_{12} = 1$. In this case, $\mathbb{E}\Pi_1 = \frac{(a-c)^2}{9b} + \frac{1}{36\alpha b} - k - r$,

Table A.1: Possible equilibria in the model of two farmers.

Equilibria	Linkages	Information Provisions	Expected Revenues
1	$(+,-)$	$\gamma_1^*=0$ $\gamma_2^*=\alpha-\beta$	$\mathbb{E}\Pi_1=\frac{(a-c)^2}{9b}+\frac{1}{36\alpha b}-k$ $\mathbb{E}\Pi_2=\frac{(a-c)^2}{9b}+\frac{1}{36\alpha b}-r$
2	$(+,-)$	$\gamma_1^*>0$ $\gamma_2^*=\alpha-\beta-\gamma_1^*$	$\mathbb{E}\Pi_1=\frac{(a-c)^2}{9b}+\frac{1}{36\alpha b}-k-r$ $\mathbb{E}\Pi_2=\frac{(a-c)^2}{9b}+\frac{1}{36\alpha b}-r$
3	$(-,-)$	$\gamma_1^*=\frac{-2\beta+\sqrt{27\alpha^2+36\alpha\beta+13\beta^2}}{9}$ $\gamma_2^*=\gamma_1^*$	$\mathbb{E}\Pi_1=\frac{(a-c)^2}{9b}-r$ $+\frac{3}{\left(18\alpha+10\beta+4\sqrt{27\alpha^2+36\alpha\beta+13\beta^2}\right)b}$ $\mathbb{E}\Pi_2=\mathbb{E}\Pi_1$
4	$(-,-)$	$\gamma_1^*=0$ $\gamma_2^*=\frac{3\alpha^2+4\alpha\beta+1\beta^2}{6\alpha+4\beta}$	$\mathbb{E}\Pi_1=\frac{(a-c)^2}{9b}+\frac{\beta}{9(\alpha+\beta)^2b}$ $\mathbb{E}\Pi_2=\frac{(a-c)^2}{9b}+\frac{9\alpha+5\beta}{36(\alpha+\beta)(2\alpha+\beta)b}-r$
5	$(-,-)$	$\gamma_1^*=0$ $\gamma_2^*=0$	$\mathbb{E}\Pi_1=\frac{(a-c)^2}{9b}+\frac{\beta}{9(\alpha+\beta)^2b}$ $\mathbb{E}\Pi_2=\mathbb{E}\Pi_1$

and $\mathbb{E}\Pi_2=\frac{(a-c)^2}{9b}+\frac{1}{36\alpha b}$. However, if the first farmer deviates by choosing $g_{12}=0$ and $\gamma_1^{'}=\frac{3\alpha^2+4\alpha\beta+1\beta^2}{6\alpha+4\beta}$, she will receive $\mathbb{E}\Pi_1^{'}=\frac{(a-c)^2}{9b}+\frac{9\alpha+5\beta}{36(\alpha+\beta)(2\alpha+\beta)b}-r>\mathbb{E}\Pi_1$. Thus, this equilibrium is not sustained.

To see whether the proposed five classes of equilibria exist, we proceed by checking all possibilities of deviations. In the first class of equilibria, if farmer 2 chooses $\gamma_2^*=0$, she will receive $\mathbb{E}\Pi_2^{'}=\frac{(a-c)^2}{9b}+\frac{\beta}{9(\alpha+\beta)^2b}$ instead of $\mathbb{E}\Pi_2=\frac{(a-c)^2}{9b}+\frac{1}{36\alpha b}-r$. Thus, such equilibrium exists if $r\leq\frac{1}{36\alpha b}-\frac{\beta}{9(\alpha+\beta)^2b}=\frac{(\alpha-\beta)^2}{36\alpha(\alpha+\beta)^2b}$. On the other hand, if farmer 1 chooses to disconnect, she will also receive $\mathbb{E}\Pi_1^{'}=\frac{(a-c)^2}{9b}+\frac{\beta}{9(\alpha+\beta)^2b}$. Thus, we need $k\leq\frac{(\alpha-\beta)^2}{36\alpha(\alpha+\beta)^2b}$. In addition, if farmer 1 disconnects and obtains her own signal, the revenue-maximizing deviation is to choose $\gamma_1^{'}=\frac{(3\alpha-\beta)(3\alpha^2+\beta^2)}{15\alpha^2-6\alpha\beta+\beta^2}$. Consequently, farmer 1 will receive $\mathbb{E}\Pi_1^{'}=\frac{(a-c)^2}{9b}+\frac{9\alpha^2-3\alpha\beta+\beta^2}{9\alpha b(15\alpha^2-2\alpha\beta-\beta^2)}-r$, and it requires that $k<r+\frac{1}{36\alpha b}-\frac{9\alpha^2-3\alpha\beta+\beta^2}{9\alpha b(15\alpha^2-2\alpha\beta-\beta^2)}$. However, since $\frac{(\alpha-\beta)^2}{36\alpha(\alpha+\beta)^2b}+\frac{1}{36\alpha b}-\frac{9\alpha^2-3\alpha\beta+\beta^2}{9\alpha b(15\alpha^2-2\alpha\beta-\beta^2)}<0$, thus existence conditions cannot be simultaneously satisfied. Similarly, the second class of equilibria do not exist since the farmer will disconnect and change her information provision.

Consider the third class of equilibrium. If farmer 1 chooses $\gamma_1^*=0$, she is worse off. Thus, to ensure that the information acquisition cost is lower than the revenue compensation, we need $r<\frac{(3\alpha+\beta)^2}{6(\alpha+\beta)^2\left(9\alpha+7\beta+2\sqrt{27\alpha^2+36\alpha\beta+13\beta^2}\right)b}$. If farmer 1 pays for the connection, she will adjust $\gamma_1^*=\alpha-\beta-\frac{\sqrt{27\alpha^2+36\alpha\beta+13\beta^2}-2\beta}{9}$, provided $\beta<\frac{9-\sqrt{57}}{4}\alpha$. By such deviation, farmer 1 gets a lower payoff of $\mathbb{E}\Pi_1^{'}=\frac{(a-c)^2}{9b}+\frac{1}{36\alpha b}-k-r$, since $\frac{1}{36\alpha b}<\frac{3}{\left(18\alpha+10\beta+4\sqrt{27\alpha^2+36\alpha\beta+13\beta^2}\right)b}$.

Finally, in the forth class of equilibria, to ensure that farmer 1 chooses $\gamma_1^*=0$, while

$\gamma_2^* = \frac{3\alpha^2+4\alpha\beta+1\beta^2}{6\alpha+4\beta}$, we need:

$$\frac{(3\alpha+\beta)^4}{24(\alpha+\beta)^2(2\alpha+\beta)(21\alpha^2+20\alpha\beta+5\beta^2)b} < r < \frac{(3\alpha+\beta)^2}{36(\alpha+\beta)^2(2\alpha+\beta)b}. \tag{A.3}$$

When $r > \frac{(3\alpha+\beta)^2}{36(\alpha+\beta)^2(2\alpha+\beta)b}$, we have the fifth class of equilibrium. In this case, if farmer 1 chooses to connect and adjust her information provision, she is worse off if $k > \frac{(\alpha-\beta)^2}{36\alpha(\alpha+\beta)^2b} - r$, which is guaranteed since $r > \frac{(3\alpha+\beta)^2}{36(\alpha+\beta)^2(2\alpha+\beta)b}$. □

Proof of Lemma 1. Firstly, applying Lemma 2 to the case when there is a single coalition, we have

$$q_i^* = \frac{a-c}{(n+1)b} + \frac{\beta}{2b(n+1)\alpha}x_0 + \frac{\sum_{i=1}^{i=N}\gamma_i x_i}{2(n+1)\alpha b}, \forall i \in N, \tag{A.4}$$

and

$$\mathbb{E}\left[\Pi_i(\gamma_i^*, \mathbf{g_i}, q_i^*)\right] = \frac{(a-c)^2}{(n+1)^2b} + \frac{1}{4(n+1)^2\alpha b} - r\delta\{\gamma_i > 0\} - k|N_i(\mathbf{g})|, \tag{A.5}$$

as long as r is small. Notice from the above expression that: (1) The farmers always respond positively towards the public signal, i.e., $\frac{\beta}{2b(n+1)\alpha} > 0$; (2) The expected payoffs $\mathbb{E}\left[\Pi_i(\gamma_i^*, \mathbf{g_i}, q_i^*)\right]$ are independent of β; (3) The value of information $\frac{1}{4(n+1)^2\alpha b}$ diminishes when the population size increases. □

Proof of Proposition 2. We first show that a connected network cannot be sustained in equilibrium. Since an endogenously formed connected network has to be a tree (see Corollary 4), there exists a farmer i connected only to a single neighbour j. There are four cases: (1) farmer i connects with j, and $\gamma_i = 0$; (2) farmer i connects with j, and $\gamma_i > 0$; (3) farmer i is connected by j, and $\gamma_i = 0$; (4) farmer i is connected by j, and $\gamma_i > 0$.

In case 1, since the network is fully connected, there exist a farmer k, $\gamma_k > 0$. The information value of her private signal is maximized when no other farmer obtains private signal. In this case, $r < \mathbb{E}\left[\Pi_k(\gamma_k, \mathbf{g_k}, q_k^*)\right] - \mathbb{E}\left[\Pi_k(0, \mathbf{g}_k, q_k^*)\right] < \frac{(\alpha-\beta)^2}{4\alpha(n+1)^2(\alpha+\beta)^2b}$. Consider the following deviation: farmer i chooses to unlink with farmer j, and obtains her own private signal x_i. The induced information structure will be asymmetric, i.e., $X_i = \{x_0, x_i\}$, $X_j = \{x_0, x_{-i}\}$. Plugging in the equilibrium information precision $\sum_{k\neq i}\gamma_k^* = \alpha - \beta$, the best responding γ_i^* can be obtained by similar procedure as in the two-farmer model

$$\gamma_i^* = \frac{\left[(n+1)\alpha^2 + (n-1)\beta^2\right]\left[(n+1)\alpha - (n-1)\beta\right]}{(n+1)(n+3)\alpha^2 - 2(n^2-1)\alpha\beta + (n-1)^2\beta^2}, \tag{A.6}$$

and thus,

$$\begin{aligned}\mathbb{E}\left[\Pi_i(\gamma_i^*, \mathbf{g_i}, q_i^*)\right] &= \frac{(a-c)^2}{(n+1)^2b} + \frac{(1+n)^2\alpha^2 + (n-1)^2\alpha\beta + (n-1)^2\beta^2}{(n+1)^2\alpha b\left[(n+1)\alpha - (n-1)\beta\right]\left[(n+3)\alpha - (n-3)\beta\right]} - r \\ &< \frac{(a-c)^2}{(n+1)^2b} + \frac{1}{4(n+1)^2\alpha b} - k, \end{aligned} \tag{A.7}$$

so that such deviation can be prevented. However, it can be checked that this contradicts to the fact that $r < \frac{(\alpha-\beta)^2}{4\alpha(n+1)^2(\alpha+\beta)^2 b}$, for $\forall n$, as $\frac{\beta}{\alpha} \to 0$. Therefore, case 1 cannot be a equilibrium.

In case 2, the same argument yields that $r < \frac{(\alpha-\beta)^2}{4\alpha(n+1)^2(\alpha+\beta)^2 b}$; otherwise farmer i should obtain no private signal. Similarly, we cannot eliminate the deviation that farmer i disconnects with farmer j and adjusts her private information provision. To see this, suppose that farmer i disconnects and adjusts her private signal provision from γ_i^* to $\gamma_i^{'}$, where

$$\begin{aligned} \lim_{\beta/\alpha \to 0} \mathbb{E}\left[\Pi_i(\gamma_i^{'}, \mathbf{g_i}, q_i^*)\right] &= \frac{(2\alpha - \gamma_i^*)^2}{4\alpha b\left[(n+1)\alpha - n\gamma_i^*\right]\left[(n+3)\alpha - (n+1)\gamma_i^*\right]} \\ &> \frac{1}{4(n+1)^2\alpha b}, \forall n, \forall \gamma_i^* < \alpha - \beta. \end{aligned} \tag{A.8}$$

Case 3 is quickly eliminated since farmer j can save the connection cost without changing her information set. In case 4, suppose that farmer j disconnects and adjusts her private signal provision to be $\gamma_j^{'}$, where

$$\begin{aligned} \lim_{\beta/\alpha \to 0} \mathbb{E}\left[\Pi_j(\gamma_j^{'}, \mathbf{g}_j, q_j^*)\right] &= \frac{(\alpha + \gamma_i^*)^2}{4\alpha b\left[n\alpha + (n+1)\gamma_i^*\right]\left[\alpha + 2\gamma_i^*\right]} \\ &> \frac{1}{4(n+1)^2\alpha b}, \forall n, \forall \gamma_i^* < \alpha - \beta. \end{aligned} \tag{A.9}$$

Thus, farmer j could be better off to unlink with farmer i and adjust her signal. □

Proof of Corollary 3. Fix the choices of ρ_j, $\forall j \neq i$, the value of the private information is increasing for $0 < \rho_i < \rho_i^*$ while decreasing for $\rho_i > \rho_i^*$, where

$$\rho_i^* = \frac{(\alpha+\beta)\left[1 + \sum_{j\neq i} \frac{n_j \rho_j}{\alpha+\beta+\rho_j}\right]}{1 + n_i + \sum_{j\neq i} \frac{n_j \rho_j}{\alpha+\beta+\rho_j}}. \tag{A.10}$$

The results follow by checking that $\frac{\partial \rho_i^*}{\partial n_i} < 0$, $\frac{\partial \rho_i^*}{\partial n_j} > 0$, and $\frac{\partial \rho_i^*}{\partial \rho_j} > 0$. For symmetric equilibria where $n_i = n^{(m)} = \frac{n}{m}$, when there are m coalitions, $\rho_i^* = \rho_i^{(m)} = \frac{(\alpha+\beta)\left[n - 2n^{(m)} + \sqrt{(n-2n^{(m)})^2 + 4n + 4}\right]}{2(n+1)}$. Plugging in the payoff function and it is straightforward to show that farmers' revenues decrease in m. □

Proof of Corollary 4. We shall prove by contradiction. Suppose that there exists a component N_i which is not a tree. By definition, there is at least a cycle, denoted by a sequence of the vertices $v_1, v_2, \cdots, v_l \in N_i$, such that $\overline{g}_{v_1 v_2} = \cdots = \overline{g}_{v_l v_1} = 1$. We can delete any edge among $e_{v_1 v_2}$, ..., $e_{v_l v_1}$, such that the vertices $v_1, v_2, \cdots, v_l$ are still connected, and thus the information structure is still the same. From Lemma 2, we reduce the linking cost by k while the same revenue is maintained for any farmer. Thus, the farmer who saves a cost k is better off. It follows, *reductio ad absurdum*, that the equilibrium network is a forest. □

Proof of Proposition 3. An empty network corresponds to the partition such that $(N_1, N_2, ..., N_m) = (\{1\}, \{2\}, ..., \{n\})$. The equilibrium precision γ_i^* is obtained via first-order

condition to Lemma 2:

$$\gamma_i^* = \frac{\Delta - 2(1+n)(\alpha+\beta)}{2(n+1)^2} > 0, \tag{A.11}$$

for $\forall i$, while the equilibrium revenue is

$$R_i^*(\{1\},\{2\},...,\{n\}) = \frac{2}{b\Delta} - \frac{4\left[(n+1)\alpha + n\beta\right]}{b\Delta},$$

where $\Delta = n(n+1)\alpha + n(n-1)\beta + \sqrt{\begin{array}{c}(n^2+8)(n+1)^2\alpha^2 + n(n^3-2n^2+9n+8)\beta^2 \\ +2(n^4+7n^2+12n+4)\alpha\beta\end{array}}$.

If farmer i chooses to connect with farmer j, she will adjust the signal precision to be $\gamma_i^{'}$, and the corresponding revenue is $R_i^{'}(\{1\},\{2\},...,\{i,j\},...,\{n\})$. The calculation procedure is similar and we omit the algebra. Finally, it can be checked that:

$$\begin{aligned} & \lim_{n\to\infty} R_i^{'}(\{1\},\{2\},...,\{i,j\},...,\{n\}) - R_i^*(\{1\},\{2\},...,\{n\}) \\ = & \lim_{n\to\infty} \frac{n\alpha(\alpha+\beta)^5 + 2\beta^5(\alpha+2\beta)}{n^3(\alpha+\beta)\left[2\beta^2(\alpha+2\beta) + n\alpha(\alpha+\beta)^2\right]^2 b}. \end{aligned}$$

For finite n, farmer i is strictly better off by merging with farmer j, as long as the cost k is small. Thus, the empty network cannot be formed in equilibrium. □

Proof of Proposition 4. The proof is similar to the standard procedure in the two-farmer model. Since β is small, we can focus on the value of private information in the generic payoff formula in Lemma 2. The equilibrium choices of the information provisions are given by the first-order conditions: $\gamma_1^* = \sqrt{\frac{n_2+1}{(n_1+1)(n+1)}}(\alpha+\beta)$, $\gamma_2^* = \sqrt{\frac{n_1+1}{(n_2+1)(n+1)}}(\alpha+\beta)$. The equilibrium revenue of any farmer in group N_i, $i=1,2$, is given by

$$\lim_{\beta\to 0} R_i^{ts} = \frac{(a-c)^2}{(n+1)^2 b} + \frac{(n+2)^2}{4\sqrt{(n_i+1)(n_{-i}+1)(n+1)}\left(\sqrt{n_{-i}+1}+\sqrt{(n_i+1)(n+1)}\right)^2(\alpha+\beta)b}.$$

Clearly the payoffs are decreasing in the public information provision β. To see that the aggregate payoff increases in the size difference $|n_1-n_2|$, we can show that $\frac{n_1}{\left(\sqrt{n_2+1}+\sqrt{(n_1+1)(n+1)}\right)^2} + \frac{n_2}{\left(\sqrt{n_1+1}+\sqrt{(n_2+1)(n+1)}\right)^2}$ decreases in $|n_1-n_2|$.

We follow the standard procedures to check all possible deviations: (1) Any one of the two farmers, $i \in N_1$ or $j \in N_2$ chooses not to obtain any signal. This could not happen if $r < \overline{r}$, where

$$\overline{r} = \overline{k} = \min\left\{ \begin{array}{c} \frac{(n+2)^2}{4\sqrt{(n_1+1)(n_2+1)(n+1)}\left(\sqrt{n_2+1}+\sqrt{(n_1+1)(n+1)}\right)^2\alpha b}, \\ \frac{(n+2)^2}{4\sqrt{(n_1+1)(n_2+1)(n+1)}\left(\sqrt{n_1+1}+\sqrt{(n_2+1)(n+1)}\right)^2\alpha b} \end{array} \right\}.$$

(2) $\forall k \in N_1$, chooses to unlink with farmer i, and obtains her own private signal. This will not happen if $k < \overline{k} = \overline{r}$ and $r > k + \underline{r}$. Due to limited space, we omit the long expression for $\underline{r}$. (3) $\forall k \in N_2$, chooses to unlink with farmer j, and obtains her own private signal. This is symmetric with the second case. (4) $\forall k \in N_1$, chooses to link with any farmer within N_2, and adjusts her signal precision or obtains additional private signal. (5) $\forall k \in N_2$, chooses to link with any farmer within N_1, and adjusts her signal precision or obtains additional private signal. The last two cases can be eliminated if $k > \underline{k}$, where $\underline{k}(\alpha, n_1, n_2) = \frac{1}{4(1+n)^2\alpha b} - \overline{r}$. It can be checked that $\underline{k} \leq 0$, when $n \geq 15$, so that the two stars have no incentive to merge. □

Proof of Corollary 5. If $\rho_i < \infty$ as $\beta \to \infty$, the value of private information is dominated by the value of public information. In this case, $\lim_{\beta\to\infty} R_i = \frac{(a-c)^2}{(1+n)^2 b} + \frac{\beta}{b} \cdot \left[\frac{1}{\alpha+\beta+\rho_i} - \frac{\sum_{i=1}^{m} \frac{n_i}{\alpha+\beta+\rho_i}}{1+n}\right]^2$, which is quasi-concave in ρ_i. Alternatively, we should consider the case where $\rho_i \to \infty$ for some i. □

Proof of Proposition 5. The proposed network with a dominant group corresponds to a partitioning $(N^*, \{i : i \in N, i \notin N^*\})$ such that $N^* \cap \{i : i \in N, i \notin N^*\} = \emptyset$ and $\cup_{\forall i, i\in N, i\notin N^*}\{i\} \cup N^* = N$. The revenue of $i \notin N^*$ is given by

$$\lim_{\gamma_i\to 0, \gamma_j\to\infty} R_i = \frac{(a-c)^2}{(1+n)^2 b} + \frac{\beta}{(1+|N^*|)^2 b} \cdot \left[\frac{1+|N^*|}{(1+n)(\alpha+\beta)}\right]^2, \tag{A.12}$$

while the revenue of $j \in N^*$ is given by

$$\lim_{\gamma_i\to 0, \gamma_j\to\infty} R_j = \frac{(a-c)^2}{(1+n)^2 b} + \frac{\beta}{(1+|N^*|)^2 b} \cdot \left[\frac{n-|N^*|}{(1+n)(\alpha+\beta)}\right]^2. \tag{A.13}$$

Note that the value of private information $\frac{\gamma_i}{(1+|N^*|)^2(\alpha+\beta+\gamma_i)^2 b} \to 0$ either as $\gamma_i \to 0$ or $\gamma_i \to \infty$.

Similar to the procedure in Proposition 1 and Proposition 4, we can show that: (1) $\forall i \notin N^*$ will not connect with $j \in N^*$, as $|N^*| \geq \frac{n}{2} - 1$; (2) $\forall i \notin N^*$ will not choose to obtain private signal with precision $\gamma_i > 0$, and/or form links with any other farmer(s), since $r > k$; (3) $\forall i \in N^*$ will not connect with $j \notin N^*$; (4) $\forall i \in N^*$ will not disconnect with the dominant group, since $|N^*| < \frac{n-1}{2}$; (5) The dominant group will not deviate from the infinite information provision; (6) $\forall i \in N^*$ will not disconnect with the dominant group, and obtains her private signal with precision $\gamma_i > 0$, since $r > k$. We omit the algebra due to limited space. □

Proof of Corollary 6. The aggregate payoff for the dominant group architecture is given by

$$\lim_{n\to\infty} \lim_{\gamma_i\to 0, \gamma_j\to\infty} \sum_{i=1}^{i=n} R_i^{dg} = \lim_{n\to\infty} \frac{n}{(1+n)^2 b} \left[(a-c)^2 + \frac{\beta}{(\alpha+\beta)^2}\right], \tag{A.14}$$

as $|N^*| \to \frac{n}{2}$. Clearly the farmers' aggregate payoff is increasing in β as long as $\beta < \alpha$. On the other hand,

$$\lim_{\gamma_i \to 0, \gamma_j \to \infty} \sum_{i=1}^{i=n} R_i = \frac{n\,(a-c)^2}{(1+n)^2 b} + \frac{\beta}{\left(1+|N^*|\right)^2 b} \cdot \left[\frac{\left(1+|N^*|\right)^2 \left(n-|N^*|\right) + \left(n-|N^*|\right)^2 |N^*|}{(1+n)^2(\alpha+\beta)^2} \right], \tag{A.15}$$

which is maximized when $|N^*| \to \frac{n-1}{n+3}$. It can be checked that the optimal $|N^*| = 1$, and

$$\lim_{n\to\infty} \lim_{\gamma_i \to 0, \gamma_j \to \infty} \sum_{i=1}^{i=n} R_i^* = \lim_{n\to\infty} \frac{n}{(1+n)^2 b} \left[(a-c)^2 + \frac{n\beta}{4(\alpha+\beta)^2} \right].$$

Therefore, $\lim_{n\to\infty} \frac{\sum_{i=1}^{i=n} R_i^{dg}}{\sum_{i=1}^{i=n} R_i^*} = \lim_{n\to\infty} \frac{(a-c)^2 + \frac{\beta}{(\alpha+\beta)^2}}{(a-c)^2 + \frac{n\beta}{4(\alpha+\beta)^2}} = 0$. Finally, consider farmer i in the dominant group and farmer j outside the dominant group, it can be checked that

$$\lim_{n\to\infty} R_i^{dg} = \lim_{n\to\infty} R_j^{dg} = \lim_{n\to\infty} \frac{(a-c)^2}{(1+n)^2 b} + \frac{\beta}{(1+n)^2(\alpha+\beta)^2 b}, \tag{A.16}$$

and the result follows.

To see the uniqueness of this class of equilibria, we need to show the following: (1) A farmer either chooses $\gamma_i \to 0$, or $\gamma_i \to \infty$, according to **Corollary 5**. Those who prefer $\gamma_i \to 0$ will be isolated since $k > 0$; (2) There is only one farmer i who obtains private signal and those who prefer high private information provision are connected with i in a tree graph N^*, since $r > k > 0$; (3) $\forall j \neq i$, $j \in N^*$, either $g_{ji} = 1$, or $g_{jv_1} = g_{v_1 v_2} = \cdots = g_{v_l i} = 1$, for some sequence of the vertices $v_1, v_2, \cdots, v_l \in N^*$. Because if there is some j such that either $g_{ij} = 1$, or $g_{vj} = 1$, for some v that is connected with i, the farmer i or v will be better off by disconnecting with j (and the subgraph it possibly connects) due to congestion effect. (4) $\frac{n}{2} - 1 \leq |N^*| < \frac{n}{2} - \frac{1}{2}$ contains only one integer. □

Proof of Proposition 6. Suppose the unit production costs are c_H and c_L respectively. Following similar procedure, we summarize the equilibrium quantities in Table A.2.

As the procedure of checking possible deviations coincides with that in the basic model, we make the same conclusion that there is no information sharing in any equilibrium. □

Proof of Proposition 7. We follow the proof of the Proposition 5 by considering the partitionings over the set N. In particular, the isolated network is a special case, such that $(N_1, N_2, ..., N_m) = (\{1\}, \{2\}, ..., \{n\})$. When $\frac{\gamma_i}{\beta} \to \infty$, we focus on the value of private information. Consider farmer i's one-link severance that corresponds to a *refinement* of the partition $(N_1, ..., N_i, ..., N_j, ..., N_m)$ while the original partition is

$$(N_1, ..., N_{i-1}, N_{i+1}, ..., N_{j-1}, N_{j+1}, ..., N_m, N_i \cap N_j).$$

We shall show that such a refinement increases farmer i's revenue, i.e.,

$$\frac{\rho_i}{\left[1 + \sum_{k=1}^{m} \left(\frac{n_k \rho_k}{\alpha+\beta+\rho_k}\right)\right]^2 (\alpha+\beta+\rho_i)^2 b} > \frac{\rho_i + \rho_j}{\left[1 + \sum_{k\neq i,j} \left(\frac{n_k \rho_k}{\alpha+\beta+\rho_k}\right) + \frac{(n_i+n_j)(\rho_i+\rho_j)}{\alpha+\beta+\rho_i+\rho_j}\right]^2 (\alpha+\beta+\rho_i+\rho_j)^2 b}.$$

Table A.2: Equilibrium characterization for the model of two farmers with heterogenous production costs.

r	Signal precisions	Expected revenues
small	$\gamma_2^* = \gamma_1^* = \frac{-2\beta+\sqrt{27\alpha^2+36\alpha\beta+13\beta^2}}{9}$	$E\Pi_1 = \frac{(a-2c_H+c_L)^2}{9b} + \frac{3}{\left(18\alpha+10\beta+4\sqrt{27\alpha^2+36\alpha\beta+13\beta^2}\right)b} - r$ $E\Pi_2 = \frac{(a-2c_L+c_H)^2}{9b} + \frac{3}{\left(18\alpha+10\beta+4\sqrt{27\alpha^2+36\alpha\beta+13\beta^2}\right)b} - r$
medium	$\forall i = 1,2, \gamma_i^* = 0,$ $\gamma_{3-i}^* = \frac{3\alpha^2+4\alpha\beta+\beta^2}{6\alpha+4\beta}$	$E\Pi_i = \frac{(a-2c_H+c_L)^2}{9b} + \frac{\beta}{9(\alpha+\beta)^2 b}$ $E\Pi_{3-i} = \frac{(a-2c_L+c_H)^2}{9b} + \frac{9\alpha+5\beta}{36(\alpha+\beta)(2\alpha+\beta)b} - r$
large	$\gamma_2^* = \gamma_1^* = 0$	$E\Pi_1 = \frac{(a-2c_H+c_L)^2}{9b} + \frac{\beta}{9(\alpha+\beta)^2 b}$ $E\Pi_2 = \frac{(a-2c_L+c_H)^2}{9b} + \frac{\beta}{9(\alpha+\beta)^2 b}$

This is because that: (1) $\frac{n_i\rho_i}{\alpha+\beta+\rho_i} + \frac{n_j\rho_j}{\alpha+\beta+\rho_j} < \frac{(n_i+n_j)(\rho_i+\rho_j)}{\alpha+\beta+\rho_i+\rho_j}$; (2) $\frac{\rho_i+\rho_j}{(\alpha+\beta+\rho_i+\rho_j)^2} - \frac{\rho_i}{(\alpha+\beta+\rho_i)^2} = \frac{\left[(\alpha+\beta)^2-(\rho_i+\rho_j)\right]\rho_j}{(\alpha+\beta+\rho_i+\rho_j)^2(\alpha+\beta+\rho_i)^2} < 0$. Thus, we can repeat such refinement to show that an empty network is the unique equilibrium.

Similarly, consider the farmers' aggregate payoff $\sum_{i=1}^m n_i \mathbb{E}\Pi_i = \sum_{i=1}^m n_i R_i\left(N_1, N_2, ..., N_m\right) - k \cdot \sum_{i=1}^n |N_i(\mathbf{g})|$, and notice that: (1) $\sum_{i=1}^m n_i$ is not affected by the graph operation; (2) $\frac{n_i\rho_i}{\alpha+\beta+\rho_i} + \frac{n_j\rho_j}{\alpha+\beta+\rho_j} < \frac{(n_i+n_j)(\rho_i+\rho_j)}{\alpha+\beta+\rho_i+\rho_j}$, which means that $\sum_{i=1}^m \left(\frac{n_i\rho_i}{\alpha+\beta+\rho_i}\right)$ will increase; (3) $\sum_{i=1}^m \left[\frac{n_i\rho_i}{(\alpha+\beta+\rho_i)^2}\right]$ will decrease. Thus, any *refinement* of the partitioning will increase the farmers' aggregate payoff due to the increase in revenue and the decrease in linking cost. This implies that the social isolation is maximizing the farmers' aggregate payoff. □

Proof of Proposition 8. Suppose that all farmers join the FPO. To see a contradiction, suppose that farmer i disconnects and adjusts her private signal provision from γ_i^* to $\gamma_i^{'}$. She receives revenue $\frac{(2\alpha-\gamma_i^*)^2}{4\alpha b\left[(n+1)\alpha-n\gamma_i^*\right]\left[(n+3)\alpha-(n+1)\gamma_i^*\right]}$, which will be strictly greater than $\frac{1}{4(n+1)^2\alpha b}$ (the revenue if she stays in the FPO). Thus, this is not sustained as Nash equilibrium.

Complete isolation is possible because FPO requires more than one farmer to improve the participants' revenue, as $k > 0$. However, consider two farmers i and j, who jointly deviate from the *status quo* by joining the FPO together and update their private information provisions. Similar to **Proposition 3**, it can be checked that:

$$\begin{aligned} & \lim_{n\to\infty} R_i^{'}\left(\{1\},\{2\},...,\{i,j\},...,\{n\}\right) - R_i^*\left(\{1\},\{2\},...,\{n\}\right) \\ = & \lim_{n\to\infty} \frac{n\alpha(\alpha+\beta)^5 + 2\beta^5(\alpha+2\beta)}{n^3(\alpha+\beta)\left[2\beta^2(\alpha+2\beta)+n\alpha(\alpha+\beta)^2\right]^2 b}. \end{aligned}$$

For finite n large enough, such deviations are strictly beneficial for both i and j, as long as the cost k is small. □

Proof of Proposition 9. Suppose that r is high that no additional information acquisition is admitted. For the bottom-up approach, we first show that farmer 1 has incentive

to disconnect with farmer 3. Given that farmer 2 choose γ_2, farmer 1 receives a payoff Π_1 if $g_{13} = 1$, and $\Pi_1^{'}$ if $g_{13} = 0$. It can be checked that for any $\gamma_2 > 0$,

$$\lim_{\beta \to 0} \left(\Pi_1^{'} - \Pi_1 \right) = k + \left[\frac{1}{\left(1 + \frac{2\gamma_2}{\alpha+\gamma_2} + \frac{\gamma_4}{\alpha}\right)^2} - \frac{1}{\left(1 + \frac{3\gamma_2}{\alpha+\gamma_2} + \frac{\gamma_4}{\alpha}\right)^2} \right] \frac{\gamma_2}{(\alpha+\gamma_2)^2 b} > 0. \quad \text{(A.17)}$$

Thus, such network is not sustained in equilibrium since γ_2 is arbitrary. For the top-down approach, however, no farmer will disconnect if k is small enough due to the value of information. □

Proof of Proposition 10. The proof here is similar to the proof of Proposition **5,** as the entire class of the dominant group equilibria correspond to the same partition here. The only possible equilibrium corresponds to the dominant group architecture: a set of farmers $N^* \subseteq N$ ($\frac{n}{2} - 1 \leq |N^*| < \frac{n}{2} - \frac{1}{2}$) join the same FPO, one of which chooses signal precision $\gamma_i^* \to \infty$, while the rest of the farmers are isolated and $\forall j \notin N^*$, $\gamma_j^* = 0$.

However, the private information holder i is better off by leaving the FPO. If she stays, she receives

$$R_i = \frac{(a-c)^2}{(1+n)^2 b} + \frac{\beta}{(1+|N^*|)^2 b} \cdot \left[\frac{1+|N^*|}{(1+n)(\alpha+\beta)} \right]^2, \quad \text{(A.18)}$$

while if she leaves, she is getting

$$R_i^{'} = \frac{(a-c)^2}{(1+n)^2 b} + \frac{\beta}{4b} \cdot \left[\frac{n-1}{(1+n)(\alpha+\beta)} \right]^2. \quad \text{(A.19)}$$

If the dominant group is Nash equilibrium, it must be that $R_i^{'} < R_i \Rightarrow n < 3$. By contradiction, there is no pure strategy Nash equilibrium. □

Proof of Proposition 11. The analysis is similar with the basic model. There are two possible equilibria where information is shared. (1) Both farmers pay $k > 0$ and $r > 0$. Consequently, $\gamma_1^* + \gamma_2^* = \alpha - \beta$, $\gamma_1^* \cdot \gamma_2^* > 0$, and $\mathbb{E}\Pi_1 = \mathbb{E}\Pi_2 = \frac{(a-c)^2}{9b} + \frac{1}{36\alpha b} - k - r$. However, we can show that each farmer is better off without the agreement. Suppose farmer 2 opts out and adjusts γ_2^* to $\gamma_2^{'}$. It can be checked that the maximum $\mathbb{E}\Pi_2^{'} = \frac{(a-c)^2}{9b} + \frac{\left(2\alpha - \gamma_2^*\right)^2}{4\alpha b\left(3\alpha - 2\gamma_2^*\right)\left(5\alpha - 3\gamma_2^*\right)} - r$, which is strictly greater than $\frac{(a-c)^2}{9b} + \frac{1}{36\alpha b} - k - r$ for $k > 0$. (2) Both farmers pay $k > 0$, but only farmer 1 pays $r > 0$ and chooses $\gamma_1^* = \alpha - \beta$. To prevent farmer 2 from disconnection and adjusting $\gamma_2^* = 0$ to $\gamma_2^{'} > 0$, we need $k < r - \frac{7}{180\alpha b}$. However, this contradicts to the necessary condition for farmer 1 to pay $r > 0$ in the first place, i.e., $r < \frac{1}{36\alpha b} - k$.

The impossibility proof of the fully connected network is similar by considering the end-node in tree-network and its neighbor. Complete isolation is not sustained in the sense of strong Nash equilibrium because two farmers are better off by reaching a mutual agreement of information sharing, while the algebra is identical to that in the proof of Proposition **8**. □

Proof of Proposition 12. Suppose that the two farmers are separated, and choose the private information precision γ_1, γ_2, respectively. Similar to the solution procedure in the basic model, we obtain the expected payoff

$$\begin{aligned}\mathbb{E}\Pi_i &= \mathbb{E}\left(p_i - c\right)^2 - r\delta\{\gamma_i > 0\} = \left(\frac{a+c}{2-b} - c\right)^2 - r\delta\{\gamma_i > 0\} \\ &+ \frac{\left[2(\alpha+\beta) + (2+b)\gamma_j\right]^2 \gamma_i}{\left[4(\alpha+\beta)^2 + 4(\alpha+\beta)(\gamma_1+\gamma_2) + (4-b^2)\gamma_1\gamma_2\right]^2} \\ &+ \frac{4\left[2(\alpha+\beta) + b\gamma_i + 2\gamma_j\right]^2 \beta}{(2-b)^2\left[4(\alpha+\beta)^2 + 4(\alpha+\beta)(\gamma_1+\gamma_2) + (4-b^2)\gamma_1\gamma_2\right]^2}. \end{aligned} \tag{A.20}$$

As β is sufficiently large, the value of the private information diminishes, and we can focus on the value of the public information. Since

$$\frac{\partial\left(\lim_{\beta\to\infty}\mathbb{E}\Pi_i\right)}{\partial\gamma_i} = -\frac{16\beta(\alpha+\beta+\gamma_j)\left[2(\alpha+\beta) + b\gamma_i + 2\gamma_j\right]\left[2(\alpha+\beta) + (2+b)\gamma_j\right]}{(2-b)\left[4(\alpha+\beta)(\alpha+\beta+\gamma_i) + 4(\alpha+\beta)\gamma_j + (4-b^2)\gamma_i\gamma_j\right]^3} < 0, \tag{A.21}$$

thus, the farmers will choose $\gamma_i = 0$ even if the cost r is small. Therefore, $\forall k > 0$ deters the farmer from connection.

When β is sufficiently small, we can focus on the value of the private information. $\lim_{\beta\to 0}\mathbb{E}\Pi_i$ is first increasing and then decreasing in γ_i, and is maximized when $\gamma_1^* = \gamma_2^* = \frac{2(\alpha+\beta)}{\sqrt{4-b^2}}$. The expected payoff is

$$\lim_{\beta\to 0}\mathbb{E}\Pi_i(\gamma_i^*, g_{ij} = 0, p_i^*) = \left(\frac{a+c}{2-b} - c\right)^2 + \left[4(2-b)(\alpha+\beta) + \frac{8(\alpha+\beta)\sqrt{2-b}}{\sqrt{2+b}}\right]^{-1}. \tag{A.22}$$

When the farmers are connected,

$$\lim_{\beta\to 0}\mathbb{E}\Pi_i(\gamma_i^*, g_{ij} = 1, p_i^*) = \left(\frac{a+c}{2-b} - c\right)^2 + \frac{1}{4\alpha\left(2-b\right)^2} - k. \tag{A.23}$$

It can be checked that

$$\lim_{\beta\to 0}\mathbb{E}\Pi_i(\gamma_i^*, g_{ij} = 1, p_i^*) - \lim_{\beta\to 0}\mathbb{E}\Pi_i(\gamma_i^*, g_{ij} = 0, p_i^*) > 0, \tag{A.24}$$

for $\forall b \in (0, 2)$. Thus, it is strictly better off for the farmers to stay connected, as long as the costs k, r are small enough. Finally, since both farmers' payoffs are maximized and aligned, the farmers' aggregate payoff is also maximized. □

Appendix B

Appendix for Chapter 3

In this appendix, we provide detailed proofs of the main results in Chapter 3. To streamline our analysis, we begin with a general proof for **Proposition 17.**

Proof of Proposition 17. A consumer in period 0 calculates the expected value-to-go according to

$$\mathrm{E}\left[u_0(a_0=0)\right] = \int_{\frac{P_1}{\hat{\beta}\delta V}}^{1} \max\left\{\beta\delta^2\theta V - \beta\delta P_1, 0\right\} dF(\theta), \tag{B.1}$$

which can be rewritten via integration by parts

$$= \beta\delta^2 V - \frac{\beta}{\hat{\beta}}\delta P_1 F\left(\frac{P_1}{\hat{\beta}\delta V}\right) - \beta\delta^2 V \int_{\frac{P_1}{\hat{\beta}\delta V}}^{1} F(\theta)d\theta - \beta\delta P_1 \bar{F}\left(\frac{P_1}{\hat{\beta}\delta V}\right). \tag{B.2}$$

In the first case, the prices satisfy

$$\mathrm{E}\left[u_0(a_0=1)\right] \geq \mathrm{E}\left[u_0(a_0=0)\right]|_{\beta^*=1}, \tag{B.3}$$

which implies that

$$P_0 \leq \beta\delta^2 V \mathrm{E}(\theta) - \left[\beta\delta^2 V - \beta\delta P_1 - \beta\delta^2 V \int_{\frac{P_1}{\delta V}}^{1} F(\theta)d\theta\right]. \tag{B.4}$$

Thus, the seller maximizes his revenue when equality holds

$$\pi = \max_{P_1} \Lambda_0 \left[\beta\delta^2 V \mathrm{E}(\theta) - \beta\delta^2 V + \beta\delta P_1 + \beta\delta^2 V \int_{\frac{P_1}{\delta V}}^{1} F(\theta)d\theta\right] + \alpha\Lambda_1 P_1 \bar{F}\left(\frac{P_1}{\beta\delta V}\right). \tag{B.5}$$

The first-order condition gives

$$H\left(\frac{P_1}{\beta\delta V}\right) = \frac{\Lambda_0\beta^2\delta^2 V}{\alpha\Lambda_1 P_1}\cdot \bar{F}\left(\frac{P_1}{\delta V}\right)\Big/ \bar{F}\left(\frac{P_1}{\beta\delta V}\right) + \frac{\beta\delta V}{P_1}, \tag{B.6}$$

while the second-order condition requires that

$$-\frac{\Lambda_0\beta}{V}f(\frac{P_1}{\delta V}) - \frac{2\alpha\Lambda_1}{\beta\delta V}f\left(\frac{P_1}{\beta\delta V}\right) - \frac{\alpha\Lambda_1 P_1}{(\beta\delta V)^2}f'\left(\frac{P_1}{\beta\delta V}\right) < 0. \tag{B.7}$$

The part $\frac{2\alpha\Lambda_1}{\beta\delta V}f\left(\frac{P_1}{\beta\delta V}\right) + \frac{\alpha\Lambda_1 P_1}{(\beta\delta V)^2}f'\left(\frac{P_1}{\beta\delta V}\right) \geq 0$ is ensured by the fact that

$$\frac{2\beta\delta V}{P_1}f\left(\frac{P_1}{\beta\delta V}\right) - f'\left(\frac{P_1}{\beta\delta V}\right) \geq 0, \tag{B.8}$$

since the function $x^2 f(x)$ is non-decreasing for all x. In the first case, the aggregate product adoption is given by

$$Q = \Lambda_0 + \Lambda_1\left[\frac{P_1}{\beta\delta V}f\left(\frac{P_1}{\beta\delta V}\right) - \frac{\beta\delta\Lambda_0 P_1}{\alpha\Lambda_1}\bar{F}\left(\frac{P_1}{\delta V}\right)\right]. \tag{B.9}$$

The second case requires that

$$\mathrm{E}\left[u_0(a_0=1)\right] < \mathrm{E}\left[u_0(a_0=0)\right]\big|_{\beta^*=\beta}, \tag{B.10}$$

which implies that

$$P_0 > \beta\delta^2 V\mathrm{E}(\theta) - \left[\beta\delta^2 V - \delta P_1 F\left(\frac{P_1}{\beta\delta V}\right) - \beta\delta^2 V\int\limits_{\frac{P_1}{\beta\delta V}}^{1} F(\theta)d\theta - \beta\delta P_1\bar{F}\left(\frac{P_1}{\beta\delta V}\right)\right]. \tag{B.11}$$

Thus, the seller's problem is

$$\max_{P_1}\alpha\left(\Lambda_0+\Lambda_1\right)P_1\bar{F}(\frac{P_1}{\beta\delta V}), \tag{B.12}$$

subject to

$$P_0 > \beta\delta^2 V\mathrm{E}(\theta) - \left[\beta\delta^2 V - \delta P_1 - \beta\delta P_1\bar{F}(\frac{P_1}{\beta\delta V}) - \beta\delta^2 V\int\limits_{\frac{P_1}{\beta\delta V}}^{1} F(\theta)d\theta\right]. \tag{B.13}$$

First-order condition gives

$$P_1^* = \beta\delta V H^{-1}\left(\frac{\beta\delta V}{P_1^*}\right), \tag{B.14}$$

where $H(\cdot) = \frac{f(\cdot)}{\bar{F}(\cdot)}$ is the hazard rate function. In this case, the aggregate product adoption is given by

$$Q = (\Lambda_0 + \Lambda_1) \frac{P_1}{\beta\delta V} f(\frac{P_1}{\beta\delta V}). \tag{B.15}$$

For the separating equilibrium, the seller's problem is

$$\max_{P_0,P_1} G(\beta^*)\Lambda_0 P_0 + \alpha \left[\bar{G}(\beta^*)\Lambda_0 + \Lambda_1\right] P_1 \bar{F}(\frac{P_1}{\beta\delta V}), \tag{B.16}$$

subject to

$$P_0 > \beta\delta^2 V \mathrm{E}(\theta) - \left[\beta\delta^2 V - \beta\delta P_1 - \beta\delta^2 V \int_{\frac{P_1}{\delta V}}^{1} F(\theta)d\theta\right], \tag{B.17}$$

$$P_0 \leq \beta\delta^2 V \mathrm{E}(\theta) - \left[\beta\delta^2 V - \delta P_1 - \beta\delta P_1 \bar{F}(\frac{P_1}{\beta\delta V}) - \beta\delta^2 V \int_{\frac{P_1}{\beta\delta V}}^{1} F(\theta)d\theta\right], \tag{B.18}$$

$$P_0 = \beta\delta^2 V \mathrm{E}(\theta) - \beta\delta^2 V + \frac{\beta}{\beta^*}\delta P_1 F\left(\frac{P_1}{\beta^*\delta V}\right) + \beta\delta^2 V \int_{\frac{P_1}{\beta^*\delta V}}^{1} F(\theta)d\theta + \beta\delta P_1 \bar{F}\left(\frac{P_1}{\beta^*\delta V}\right) \tag{B.19}$$

Suppose that an interior solution β^* exists, the seller's problem becomes

$$\max_{\beta^*,P_1} G(\beta^*)\Lambda_0 \left[\begin{array}{l} \beta\delta^2 V \mathrm{E}(\theta) - \beta\delta^2 V + \frac{\beta}{\beta^*}\delta P_1 F\left(\frac{P_1}{\beta^*\delta V}\right) \\ +\beta\delta^2 V \int_{\frac{P_1}{\beta^*\delta V}}^{1} F(\theta)d\theta + \beta\delta P_1 \bar{F}\left(\frac{P_1}{\beta^*\delta V}\right) \end{array}\right] + \alpha \left[\bar{G}(\beta^*)\Lambda_0 + \Lambda_1\right] P_1 \bar{F}\left(\frac{P_1}{\beta\delta V}\right), \tag{B.20}$$

and the first-order condition implies that

$$\begin{array}{r} G(\beta^*)\Lambda_0 \left[\left(\frac{1}{\beta^*} - 1\right) \frac{\beta P_1}{\beta^* V} f\left(\frac{P_1}{\beta^*\delta V}\right) + \beta\delta \bar{F}\left(\frac{P_1}{\beta^*\delta V}\right)\right] \\ +\alpha \left[\bar{G}(\beta^*)\Lambda_0 + \Lambda_1\right] \left[\bar{F}\left(\frac{P_1}{\beta\delta V}\right) - \frac{P_1}{\beta\delta V} f\left(\frac{P_1}{\beta\delta V}\right)\right] \end{array} = 0, \tag{B.21}$$

which gives the result

$$H\left(\frac{P_1}{\beta\delta V}\right) = \frac{G(\beta^*)\Lambda_0 \left[\left(\frac{1}{\beta^*} - 1\right) \frac{\beta^2\delta}{\beta^*} f\left(\frac{P_1}{\beta^*\delta V}\right) + \frac{\beta^2\delta^2 V}{P_1} \bar{F}\left(\frac{P_1}{\beta^*\delta V}\right)\right]}{\alpha \left[\bar{G}(\beta^*)\Lambda_0 + \Lambda_1\right] \bar{F}\left(\frac{P_1}{\beta\delta V}\right)} + \frac{\beta\delta V}{P_1}. \tag{B.22}$$

In this case, the aggregate product adoption is given by

$$Q = G(\beta^*)\Lambda_0 + \left[\bar{G}(\beta^*)\Lambda_0 + \Lambda_1\right] \left\{ \begin{array}{c} \frac{P_1}{\beta\delta V} f\left(\frac{P_1}{\beta\delta V}\right) - \frac{G(\beta^*)\Lambda_0}{\alpha\left[\bar{G}(\beta^*)\Lambda_0+\Lambda_1\right]} \cdot \\ \left[\left(\frac{1}{\beta^*}-1\right)\frac{\beta P_1}{\beta^* V} f\left(\frac{P_1}{\beta^*\delta V}\right) + \beta\delta\bar{F}\left(\frac{P_1}{\beta^*\delta V}\right)\right] \end{array} \right\}. \quad \text{(B.23)}$$

To summarize, the spot-period prices in equilibrium-D, -N, and -P are the solutions to the following equations:

$$P_1^D = \max\left\{\beta\delta V H^{-1}\left[\frac{\Lambda_0\beta^2\delta^2 V}{\alpha\Lambda_1 P_1^D}\cdot\bar{F}\left(\frac{P_1^D}{\delta V}\right)\Big/\bar{F}\left(\frac{P_1^D}{\beta\delta V}\right) + \frac{\beta\delta V}{P_1^D}\right], \beta\delta V\right\}, \quad \text{(B.24)}$$

$$P_1^N = \max\left\{\beta\delta V H^{-1}\left(\frac{\beta\delta V}{P_1^N}\right), \beta\delta V\right\}. \quad \text{(B.25)}$$

$$P_1^P = \max\left\{\beta\delta V H^{-1}\left\{\frac{G(\beta^*)\Lambda_0\left[\left(\frac{1}{\beta^*}-1\right)\frac{\beta^2\delta}{\beta^*} f\left(\frac{P_1^P}{\beta^*\delta V}\right) + \frac{\beta^2\delta^2 V}{P_1^P}\bar{F}\left(\frac{P_1^P}{\beta^*\delta V}\right)\right]}{\alpha\left[\bar{G}(\beta^*)\Lambda_0+\Lambda_1\right]\bar{F}\left(\frac{P_1^P}{\beta\delta V}\right)} + \frac{\beta\delta V}{P_1^P}\right\}, \beta\delta V\right\}. \quad \text{(B.26)}$$

For comparative statics, we first show that $P_1^N < P_1^D$ by contradictions. Suppose that $P_1^N \geq P_1^D$. Since they are both interior solutions, we have

$$\begin{aligned} P_1^N &= \beta\delta V H^{-1}\left[\frac{\beta\delta V}{P_1^N}\right] \leq \beta\delta V H^{-1}\left[\frac{\beta\delta V}{P_1^D}\right] \\ &< \beta\delta V H^{-1}\left[\frac{\beta\delta V}{P_1^D} + \Delta^D\right] = P_1^D, \end{aligned} \quad \text{(B.27)}$$

where the second inequality is due to the fact that the hazard rate function $H(\cdot)$ is non-decreasing and thus, so does $H^{-1}(\cdot)$, while in the third strict inequality $\Delta^D = \frac{\Lambda_0\beta^2\delta^2 V}{\alpha\Lambda_1 P_1^D}\cdot\bar{F}\left(\frac{P_1^D}{\delta V}\right)\Big/\bar{F}\left(\frac{P_1^D}{\beta\delta V}\right) > 0$. *Reductio ad absurdum*, it must be that $P_1^N < P_1^D$. Similarly, we can show that $P_1^N < P_1^P$ by noticing that $\Delta^P = \frac{G(\beta^*)\Lambda_0\left[\left(\frac{1}{\beta^*}-1\right)\frac{\beta^2\delta}{\beta^*} f\left(\frac{P_1^S}{\beta^*\delta V}\right) + \frac{\beta^2\delta^2 V}{P_1^S}\bar{F}\left(\frac{P_1^P}{\beta^*\delta V}\right)\right]}{\alpha\left[\bar{G}(\beta^*)\Lambda_0+\Lambda_1\right]\bar{F}\left(\frac{P_1^P}{\beta\delta V}\right)} > 0$. As $\beta^* \to 0$, $\Delta^P \to 0$, and thus, $\lim_{\beta^*\to 0} P_1^P = P_1^N$. As $\beta^* \to 1$, $\Delta^P \to \Delta^D$, and thus, $\lim_{\beta^*\to 1} P_1^P = P_1^D$.

Compared with the Q^N, the adoption rate among Λ_1 decreases in the separating equilibrium, because $\Lambda_1\bar{F}\left(\frac{P_1^P}{\beta\delta V}\right) < \Lambda_1\bar{F}\left(\frac{P_1^N}{\beta\delta V}\right)$. Compared with the Q^D, the adoption rate among Λ_0 decreases in the separating equilibrium, because $G(\beta^*)\Lambda_0 + \bar{G}(\beta^*)\Lambda_0\bar{F}\left(\frac{P_1^P}{\beta\delta V}\right) < \Lambda_0$. Similar component-wise comparison implies that, Q^D increases in the the adoption rate among Λ_0 and decreases in the the adoption rate among Λ_1, when compared with Q^N. □

Proof of Proposition 13. The results for equilibrium-D and -N follows directly from the general proof for **Proposition 17.** We omit the algebra due to limited space.

For the separating equilibrium-P, the seller's problem is

$$\max_{P_0,P_1} \gamma\Lambda_0 P_0 + \alpha\left[(1-\gamma)\Lambda_0+\Lambda_1\right]P_1\left(1-\frac{P_1}{\beta\delta V}\right), \tag{B.28}$$

subject to

$$\frac{\beta\delta^2 V}{2} - \int_{\frac{P_1}{\delta V}}^{1}\left(\beta\delta^2\theta V-\beta\delta P_1\right)d\theta < P_0 \leq \frac{\beta\delta^2 V}{2} - \int_{\frac{P_1}{\beta\delta V}}^{1}\left(\beta\delta^2\theta V-\beta\delta P_1\right)d\theta. \tag{B.29}$$

It can be checked that the second constraint is binding. Plugging it back into the objective function, the first-order condition gives

$$P_1^P = \frac{\beta\delta\left\{\alpha\left[(1-\gamma)\Lambda_0+\Lambda_1\right]+\gamma\beta\delta\Lambda_0\right\}V}{2\alpha\left[(1-\gamma)\Lambda_0+\Lambda_1\right]-\gamma\left(1-2\beta\right)\delta\Lambda_0}.$$

$$P_1^P > 0 \Longrightarrow \beta > \frac{1}{2} - \frac{\alpha\left[(1-\gamma)\Lambda_0+\Lambda_1\right]}{\gamma\delta\Lambda_0}. \tag{B.30}$$

$$P_1^P < \beta\delta V \Longrightarrow \beta > 1 - \frac{\alpha\left[(1-\gamma)\Lambda_0+\Lambda_1\right]}{\gamma\delta\Lambda_0}.$$

If $\beta \geq 1-\frac{\alpha[(1-\gamma)\Lambda_0+\Lambda_1]}{\gamma\delta\Lambda_0}$, then $P_1^P \in (0,\beta\delta V)$, and we are guaranteed an interior solution. If $\frac{1}{2}-\frac{\alpha[(1-\gamma)\Lambda_0+\Lambda_1]}{\gamma\delta\Lambda_0} < \beta < 1-\frac{\alpha[(1-\gamma)\Lambda_0+\Lambda_1]}{\gamma\delta\Lambda_0}$, then $P_1^P=\beta\delta V$, $P_0^P=\frac{\beta\delta^2 V}{2}$, $\pi^P=\frac{\gamma\beta\delta^2\Lambda_0 V}{2}$. If $\beta \leq \frac{1}{2}-\frac{\alpha[(1-\gamma)\Lambda_0+\Lambda_1]}{\gamma\delta\Lambda_0}$, then $P_1^P=0$, $P_0^P=0$, $\pi^P=0$. □

Proof of Corollary 7. Firstly, $Q^D > Q^N$, if

$$\Lambda_0+\Lambda_1\left[\frac{1}{2}-\frac{\beta(2-\beta)\delta\Lambda_0}{4\alpha\Lambda_1+2\beta^2\delta\Lambda_0}\right] > \frac{\Lambda_0+\Lambda_1}{2}, \tag{B.31}$$

which requires $\beta^2\delta\Lambda_0 > \left[\beta(2-\beta)\delta-2\alpha\right]\Lambda_1$. This is always true, since $\beta(2-\beta)\delta-2\alpha < \delta-2\alpha<0$.

Secondly, if either $\gamma<\frac{1}{2}$ or $\frac{\delta}{\alpha}<\frac{2(1-\gamma)}{(2\gamma-1)\gamma}$, we have $Q^D>Q^P$ by checking that

$$\Lambda_0+\Lambda_1\left[\frac{1}{2}-\frac{\beta(2-\beta)\delta\Lambda_0}{4\alpha\Lambda_1+2\beta^2\delta\Lambda_0}\right] > \gamma\Lambda_0+\left[(1-\gamma)\Lambda_0+\Lambda_1\right]\frac{\left[\alpha(1-\gamma)-\gamma\left(1-\beta\right)\delta\right]\Lambda_0+\alpha\Lambda_1}{2\alpha\left[(1-\gamma)\Lambda_0+\Lambda_1\right]-\gamma\left(1-2\beta\right)\delta\Lambda_0}.$$

Finally, if $\left[\delta-2\alpha(1-\gamma)\right]\Lambda_0<0$, we have $Q^P>Q^N$ by checking that

$$\gamma\Lambda_0+\left[(1-\gamma)\Lambda_0+\Lambda_1\right]\left[\frac{1}{2}-\frac{\gamma\delta\Lambda_0}{4\alpha\left[(1-\gamma)\Lambda_0+\Lambda_1\right]-2\gamma\left(1-2\beta\right)\delta\Lambda_0}\right] > \frac{\Lambda_0+\Lambda_1}{2}.$$

□

Proof of Corollary 8. $Q^D = \Lambda_0 + \Lambda_1 \left[\frac{1}{2} - \frac{\beta(2-\beta)\delta\Lambda_0}{4\alpha\Lambda_1+2\beta^2\delta\Lambda_0}\right]$, it can be checked that $\frac{\partial Q^D}{\partial \delta} = -\frac{\alpha\beta(2-\beta)\Lambda_0\Lambda_1}{(2\alpha\Lambda_1+\beta^2\delta\Lambda_0)^2} < 0$. $Q^N = \frac{\Lambda_0+\Lambda_1}{2}$, which is a constant regardless of δ. If $\beta \geq 1 - \frac{\alpha[(1-\gamma)\Lambda_0+\Lambda_1]}{\gamma\delta\Lambda_0}$, $Q^P = \gamma\Lambda_0 + [(1-\gamma)\Lambda_0 + \Lambda_1]\left[\frac{1}{2} - \frac{\gamma\delta\Lambda_0}{4\alpha[(1-\gamma)\Lambda_0+\Lambda_1]-2\gamma(1-2\beta)\delta\Lambda_0}\right]$, $\frac{\partial Q^P}{\partial \delta} = -\frac{\alpha\gamma\Lambda_0[(1-\gamma)\Lambda_0+\Lambda_1]}{\{2\alpha[(1-\gamma)\Lambda_0+\Lambda_1]-\gamma(1-2\beta)\delta\Lambda_0\}^2} < 0$. If $\beta < 1 - \frac{\alpha[(1-\gamma)\Lambda_0+\Lambda_1]}{\gamma\delta\Lambda_0}$, then the adoption rate is again independent of δ. □

Proof of Corollary 9. Q^P increases in β as the denominator $\frac{\gamma\delta\Lambda_0}{4\alpha[(1-\gamma)\Lambda_0+\Lambda_1]-2\gamma(1-2\beta)\delta\Lambda_0}$ increases in β. Q^N is a constant independent of β.

$$\frac{\partial Q^D}{\partial \beta} = -\frac{\delta\Lambda_0\Lambda_1\left[2\alpha(1-\beta)\Lambda_1 - \beta^2\delta\Lambda_0\right]}{\left(2\alpha\Lambda_1 + \beta^2\delta\Lambda_0\right)^2} < 0, \tag{B.32}$$

when $\frac{\delta}{\alpha} < \frac{2(1-\beta)\Lambda_1}{\beta^2\Lambda_0}$. □

Proof of Proposition 14. To streamline the analysis for subsidizing the spot-period, we begin by examine the donor's objective function $(W-s)\,Q(s)$. If $Q(s)$ is linearly increasing in s for $s < W$, then $(W-s)\,Q(s)$ is maximized when s^* satisfies $\frac{\partial Q(s^*)}{\partial s} = \frac{Q(s^*)}{(W-s^*)}$. Since $\frac{Q(s)}{(W-s)}$ is increasing in s for $s < W$, we know that s^* increases if $\frac{\partial Q(s)}{\partial s}$ increases independent of s. In what follows, we focus on $\frac{\partial Q(s)}{\partial s}$, i.e., the marginal increase in product adoption for unit subsidy.

For the separating equilibrium, the seller's problem is

$$\max_{P_0,P_1} \gamma\Lambda_0 P_0 + \alpha\left[(1-\gamma)\Lambda_0 + \Lambda_1\right] P_1 \left(1 - \frac{P_1 - s}{\beta\delta V}\right), \tag{B.33}$$

subject to

$$\frac{\beta\delta^2 V}{2} - \int_{\frac{P_1-s}{\delta V}}^{1} \left[\beta\delta^2\theta V - \beta\delta\left(P_1 - s\right)\right] d\theta < P_0 \leq \frac{\beta\delta^2 V}{2} - \int_{\frac{P_1-s}{\beta\delta V}}^{1} \left[\beta\delta^2\theta V - \beta\delta\left(P_1 - s\right)\right] d\theta. \tag{B.34}$$

It can be checked that the second constraint is binding. Plugging it back into the objective function, the first-order condition gives

$$P_1^P(s) = \frac{\alpha\left[(1-\gamma)\Lambda_0 + \Lambda_1\right](\delta\beta V + s) + \delta\gamma\Lambda_0\left[\beta^2\delta V - (1-2\beta)s\right]}{2\alpha\left[(1-\gamma)\Lambda_0 + \Lambda_1\right] - \gamma\left(1-2\beta\right)\delta\Lambda_0}. \tag{B.35}$$

A sufficient condition to ensure that $P_1^P(s) > 0$, for all s, is to require that $\beta > \frac{1}{2} - \frac{\alpha[(1-\gamma)\Lambda_0+\Lambda_1]}{2\gamma\delta\Lambda_0}$.

Let $\bar{\beta}(s)$ be the threshold beyond which $P_1^P(s) \geq \beta\delta V = s$, i.e., $\bar{\beta}(s)$ is the larger root that solves

$$\beta\delta\left\{\gamma\left(1-\beta\right)\delta\Lambda_0 - \alpha\left[(1-\gamma)\Lambda_0 + \Lambda_1\right]\right\} V = \alpha\left[(1-\gamma)\Lambda_0 + \Lambda_1\right] s. \tag{B.36}$$

Notice when the right-hand side decreases to zero, $\bar{\beta}(s) \to 1 - \frac{\alpha[(1-\gamma)\Lambda_0+\Lambda_1]}{\gamma\delta\Lambda_0}$. For $s > 0$, $\bar{\beta}(s) < 1 - \frac{\alpha[(1-\gamma)\Lambda_0+\Lambda_1]}{\gamma\delta\Lambda_0}$.

If $\beta \geq \bar{\beta}(s)$, then $P_1^P \in (0, \beta\delta V)$, and we are guaranteed an interior solution $P_1^P(0) = \frac{\beta\delta\{\alpha[(1-\gamma)\Lambda_0+\Lambda_1]+\gamma\beta\delta\Lambda_0\}V}{2\alpha[(1-\gamma)\Lambda_0+\Lambda_1]-\gamma(1-2\beta)\delta\Lambda_0}$. This implies that $P_1^P(s) - s \leq P_1^P(0)$. For each dollar of subsidy, a fraction of $\frac{\alpha[(1-\gamma)\Lambda_0+\Lambda_1]}{2\alpha[(1-\gamma)\Lambda_0+\Lambda_1]-\gamma(1-2\beta)\delta\Lambda_0}$ goes to consumers, while a fraction of $\frac{\alpha[(1-\gamma)\Lambda_0+\Lambda_1]-\gamma(1-2\beta)\delta\Lambda_0}{2\alpha[(1-\gamma)\Lambda_0+\Lambda_1]-\gamma(1-2\beta)\delta\Lambda_0}$ goes to the seller. The consumers' share of the subsidy decreases in β, i.e., the degree of present-bias increases consumers' share of subsidy. Furthermore,

$$\begin{aligned} Q^P(s) &= \gamma\Lambda + [(1-\gamma)\Lambda_0 + \Lambda_1]\left\{1 - \frac{P_1^*(s) - s}{\beta\delta V}\right\} \\ &= Q^P(0) + \frac{\frac{\alpha}{\beta\delta V}\left[(1-\gamma)\Lambda_0 + \Lambda_1\right]^2 s}{2\alpha\left[(1-\gamma)\Lambda_0 + \Lambda_1\right] - \gamma\left(1-2\beta\right)\delta\Lambda_0}. \end{aligned} \tag{B.37}$$

Thus, $\frac{\partial Q^P(s)}{\partial s} = \frac{\alpha[(1-\gamma)\Lambda_0+\Lambda_1]^2}{\{2\alpha[(1-\gamma)\Lambda_0+\Lambda_1]-\gamma(1-2\beta)\delta\Lambda_0\}\beta\delta V} > 0$. Again, it is easily checked that the product adoption rate increases in the degree of present-bias, i.e., $\frac{\partial Q^P(s)}{\partial s}$ decreases in β. In addition,

$$\frac{\partial^2 Q^P(s)}{\partial\gamma\partial s} = -\frac{\alpha\Lambda_0\left[(1-\gamma)\Lambda_0 + \Lambda_1\right]\left\{2\alpha\left[(1-\gamma)\Lambda_0 + \Lambda_1\right] - \gamma\left(1-2\beta\right)\delta\left[(1+\gamma)\Lambda_0 + \Lambda_1\right]\right\}}{\beta\delta V\left\{2\alpha\left[(1-\gamma)\Lambda_0 + \Lambda_1\right] - \gamma\left(1-2\beta\right)\delta\Lambda_0\right\}^2}, \tag{B.38}$$

which means $\frac{\partial Q^P(s)}{\partial s}$ decreases in γ when $\beta > \frac{1}{2} - \frac{\alpha[(1-\gamma)\Lambda_0+\Lambda_1]}{\gamma\delta[(1+\gamma)\Lambda_0+\Lambda_1]}$, and increases in γ otherwise.

If $\frac{1}{2} - \frac{\alpha[(1-\gamma)\Lambda_0+\Lambda_1]}{2\gamma\delta\Lambda_0} < \beta < \bar{\beta}(s)$, then $P_1^P = \beta\delta V + s$, $\frac{\partial Q^P(s)}{\partial s} = 0$ since the adoption rate is independent of subsidy level.

If all consumers make purchase in period 0, $P_0 = \beta\delta^2 V\mathrm{E}(\theta) - \int\limits_{\frac{P_1-s}{\delta V}}^{1}\left[\beta\delta^2\theta V - \beta\delta\left(P_1 - s\right)\right]d\theta$.

The seller announces spot-period price to maximize $\pi = \Lambda_0 P_0 + \alpha\Lambda_1 P_1\left(1 - \frac{P_1-s}{\beta\delta V}\right)$, which gives

$$\begin{aligned} P_1^D(s) &= \frac{\alpha\Lambda_1\left(s + \beta\delta V\right) + \beta^2\delta\Lambda_0\left(s + \delta V\right)}{2\alpha\Lambda_1 + \beta^2\delta\Lambda_0} \\ &= P_1^D(0) - \frac{\alpha\Lambda_1 s}{2\alpha\Lambda_1 + \beta^2\delta\Lambda_0} + s. \end{aligned} \tag{B.39}$$

Similar to the separating equilibrium, we can see that for each unit of subsidy, a fraction of $\frac{\alpha\Lambda_1}{2\alpha\Lambda_1+\beta^2\delta\Lambda_0}$ goes to the consumers while the rest is shared by the seller. The aggregate product adoption is $Q^D(s) = \Lambda_0 + \Lambda_1\left\{\frac{\alpha\Lambda_1-\beta(1-\beta)\delta\Lambda_0}{2\alpha\Lambda_1+\beta^2\delta\Lambda_0} + \frac{\alpha\Lambda_1 s}{(2\alpha\Lambda_1+\beta^2\delta\Lambda_0)\beta\delta V}\right\}$. Thus,

$$\frac{\partial Q^D(s)}{\partial s} = \frac{\alpha\Lambda_1^2}{\left(2\alpha\Lambda_1 + \beta^2\delta\Lambda_0\right)\beta\delta V} > 0. \tag{B.40}$$

It is easily checked that $\frac{\partial Q^D(s)}{\partial s}$ decreases in β.

If no consumers make purchase in period 0, the seller announces spot-period price to maximize $\pi = \alpha(\Lambda_0 + \Lambda_1) P_1 \left(1 - \frac{P_1 - s}{\beta\delta V}\right)$, which gives $P_1^N = \frac{\beta\delta V + s}{2}$. The aggregate product adoption is $Q^N = \frac{1}{2}(\Lambda_0 + \Lambda_1)(\frac{1}{2} + \frac{s}{2\beta\delta V})$.

$$\frac{\partial Q^N(s)}{\partial s} = \frac{\Lambda_0 + \Lambda_1}{4\beta\delta V} > 0. \tag{B.41}$$

Again, $\frac{\partial Q^N(s)}{\partial s}$ decreases in β. □

Proof of Proposition 15. Firstly, if the donor subsidizes in the advance period by $s > 0$, the seller responds by increasing P_0 to $P_0 + s$. Thus, the donor subsidizes the seller without benefitting the consumers, and the product adoption remains the same for fixed pricing strategy. However, the revenues under different pricing strategies are affected differently. In the pooling equilibrium with discount pricing, $\pi^D(s) = \pi^D(0) + \Lambda_0 s$, while in the separating equilibrium, $\pi^P(s) = \pi^P(0) + \gamma\Lambda_0 s$. In the pooling equilibrium with no advance selling, $\pi^N(s) = \pi^N(0)$. Thus, if either $\gamma < \frac{1}{2}$ or $\frac{\delta}{\alpha} < \frac{2(1-\gamma)}{(2\gamma-1)\gamma}$, $Q^D > Q^P$. Suppose that $\pi^D(0) < \pi^P(0)$. There exits some threshold s^{DP} such that $\pi^D(s)$ cross $\pi^N(s)$ at s^{DP} from below. Thus, increasing s may increase the adoption rate due to the shift in pricing regime. Similar shift is true when $\frac{\delta}{\alpha} < 2(1-\gamma)$, and $\pi^P(0) < \pi^N(0)$. □

Proof of Proposition 16. The analysis is similar with that in the basic model without return policy. The only difference is in the calculation of $\mathrm{E}[u_0(a_0 = 1)]$. Since in period 2, a consumer receive θV if she consumes the product, or the compensation R if she decides to return it. Thus,

$$\mathrm{E}[u_0(a_0 = 1)] = \beta\delta^2 \mathrm{E}[\max\{\theta V, R\}] - P_0. \tag{B.42}$$

In the separating equilibrium,

$$P_0^P = \beta\delta^2 \left(\int_{\frac{R}{V}}^{1} \theta V d\theta + \int_{0}^{\frac{R}{V}} R d\theta \right) - \int_{\frac{P_1}{\beta\delta V}}^{1} \left(\beta\delta^2\theta V - \beta\delta P_1\right) d\theta. \tag{B.43}$$

Similarly, in the pooling equilibrium,

$$P_0^D = \beta\delta^2 \left(\int_{\frac{R}{V}}^{1} \theta V d\theta + \int_{0}^{\frac{R}{V}} R d\theta \right) - \int_{\frac{P_1}{\delta V}}^{1} \left(\beta\delta^2\theta V - \beta\delta P_1\right) d\theta. \tag{B.44}$$

Notice that P_0 in both equilibria is shifted by a constant. This is equivalent to a subsidy level s in the advance period, where

$$s = \beta\delta^2 \left(\int_{\frac{R}{V}}^{1} \theta V d\theta + \int_{0}^{\frac{R}{V}} R d\theta \right) - \frac{\beta\delta^2 V}{2} = \frac{\beta\delta^2 R^2}{2V}. \tag{B.45}$$

□

Proof of Proposition 18. If all consumers make purchase in period 0,

$$P_0 = \beta_L \left[\delta^2 V \mathrm{E}(\theta) - \int_{\frac{P_1}{\hat{\beta}\delta V}}^{1} \left(\delta^2 \theta V - \delta P_1 \right) d\theta \right].$$

The seller set P_1 to maximize $\pi = \Lambda_0 P_0 + \alpha \Lambda_1 P_1 \left[\left(1 - \frac{P_1}{\beta_H \delta V} \right) \rho + \left(1 - \frac{P_1}{\beta_L \delta V} \right) (1-\rho) \right]$, which gives

$$\begin{aligned} P_1^D &= \frac{\beta_L \beta_H \hat{\beta}^2 \delta (\beta_L \delta \Lambda_0 + \alpha \Lambda_1) V}{2\alpha\hat{\beta}^2 \left[\beta_L \rho \Lambda_1 + \beta_H (1-\rho) \Lambda_1 \right] - \beta_L^2 \beta_H (1 - 2\hat{\beta}) \delta \Lambda_0}, \\ \pi^D &= \frac{\beta_L \beta_H \hat{\beta}^2 \delta (\beta_L \delta \Lambda_0 + \alpha \Lambda_1)^2 V}{4\alpha\hat{\beta}^2 \left[(1-\rho)\beta_H + \rho \beta_L \right] \Lambda_1 - 2\beta_L^2 \beta_H (1 - 2\hat{\beta}) \delta \Lambda_0}. \end{aligned} \tag{B.46}$$

To ensure that an interior solution of P_1^D exists, we need

$$\frac{\beta_L \beta_H \hat{\beta}^2 \delta (\beta_L \delta \Lambda_0 + \alpha \Lambda_1) V}{2\alpha\hat{\beta}^2 \left[\beta_L \rho \Lambda_1 + \beta_H (1-\rho) \Lambda_1 \right] - \beta_L^2 \beta_H (1 - 2\hat{\beta}) \delta \Lambda_0} < \beta_L \delta V, \tag{B.47}$$

One sufficient condition is to require that

$$\frac{\delta}{\alpha} < \frac{2\hat{\beta}^2 \beta_L \rho + \hat{\beta}^2 \beta_H (1-2\rho)}{\beta_L^2 \beta_H (1-2\hat{\beta}) - \beta_H \hat{\beta}^2 \beta_L} \cdot \frac{\Lambda_1}{\Lambda_0}. \tag{B.48}$$

Notice that the right-hand side is positive if $\frac{\beta_L}{\beta_H} < 1 - \frac{\rho}{2}$ and $\hat{\beta} > \frac{1}{1+\sqrt{1+1/\beta_H}}$.

If no consumers make purchase in period 0, the seller announces spot-period price to maximize $\pi = \alpha (\Lambda_0 + \Lambda_1) P_1 \left[\left(1 - \frac{P_1}{\beta_H \delta V} \right) \rho + \left(1 - \frac{P_1}{\beta_L \delta V} \right) (1-\rho) \right]$, which gives $P_1^N = \frac{\beta_H \beta_L \delta V}{2[\beta_H (1-\rho) + \beta_L \rho]}$, $P_0^N > \beta_H \left[\delta^2 V \mathrm{E}(\theta) - \int_{\frac{P_1}{\hat{\beta}\delta V}}^{1} (\delta^2 \theta V - \delta P_1) \, d\theta \right]$. Thus, $\pi^N = \frac{\alpha \beta_L \beta_H \delta (\Lambda_0 + \Lambda_1) V}{4[(1-\rho)\beta_H + \rho \beta_L]}$. To ensure that an interior solution of P_1^N exists, we need $\frac{\beta_H \beta_L \delta V}{2[\beta_H (1-\rho) + \beta_L \rho]} < \beta_H \delta V$, which implies that either $1 - \frac{\rho}{2} < \frac{\beta_L}{\beta_H} < \frac{2(1-\rho)}{1-2\rho}$, when $\rho < \frac{1}{2}$, or $\rho \geq \frac{1}{2}$.

For the separating equilibrium, the seller's problem is to maximize $\pi = P_0 \rho \Lambda_0 + \alpha (\Lambda_0 + \Lambda_1)(1-\rho) P_1 \left(1 - \frac{P_1}{\beta_L \delta V} \right) + \alpha \Lambda_1 \rho P_1 \left(1 - \frac{P_1}{\beta_H \delta V} \right)$. It can be checked that

$$P_0^P = \beta_H \left[\delta^2 V \mathrm{E}(\theta) - \int_{\frac{P_1}{\hat{\beta}\delta V}}^{1} \left(\delta^2 \theta V - \delta P_1 \right) d\theta \right].$$

Plugging it back into the objective function, the first-order condition gives

$$\pi^P = \frac{\beta_L\beta_H\hat{\beta}^2\delta\left\{\beta_H\delta\Lambda_0\rho+\alpha\left[\Lambda_0(1-\rho)+\Lambda_1\right]\right\}^2 V}{4\alpha\hat{\beta}^2\left[\beta_H\left(\Lambda_0+\Lambda_1\right)(1-\rho)+\beta_L\Lambda_1\rho\right]-2\beta_H^2(1-2\hat{\beta})\beta_L\delta\Lambda_0\rho}, \tag{B.49}$$

where

$$P_1^P = \frac{\beta_L\beta_H\hat{\beta}^2\delta\left\{\beta_H\delta\Lambda_0\rho+\alpha\left[\Lambda_0(1-\rho)+\Lambda_1\right]\right\} V}{2\alpha\hat{\beta}^2\left[\beta_H\left(\Lambda_0+\Lambda_1\right)(1-\rho)+\beta_L\Lambda_1\rho\right]-\beta_H^2(1-2\hat{\beta})\beta_L\delta\Lambda_0\rho}. \tag{B.50}$$

$$P_1^P > 0 \Longrightarrow 2\alpha\hat{\beta}^2\beta_H\left(\Lambda_0+\Lambda_1\right)(1-\rho)+2\alpha\hat{\beta}^2\beta_L\Lambda_1\rho-\beta_H^2\beta_L\delta\Lambda_0\rho+2\hat{\beta}\beta_H^2\beta_L\delta\Lambda_0\rho>0, \tag{B.51}$$

which implies that either $\hat{\beta}>\frac{1}{2}$, or $\frac{\delta}{\alpha}<\frac{2\hat{\beta}^2\beta_H(\Lambda_0+\Lambda_1)(1-\rho)+2\hat{\beta}^2\beta_L\Lambda_1\rho}{(1-2\hat{\beta})\beta_H^2\beta_L\Lambda_0\rho}$. On the other hand,

$$P_1^P<\beta_L\delta V\Longrightarrow\frac{\delta}{\alpha}<\frac{\hat{\beta}^2\beta_H\Lambda_0(1-\rho)+\hat{\beta}^2\beta_H\Lambda_1(1-2\rho)+2\hat{\beta}^2\beta_L\Lambda_1\rho}{\beta_H^2\left(\beta_L+\hat{\beta}^2-2\hat{\beta}\beta_L\right)\Lambda_0\rho}. \tag{B.52}$$

□

Proof of Corollary 10. The adoption quantity for the separating equilibrium is

$$Q^P=\Lambda_0\rho+\left(1-\frac{P_1^P}{\beta_H\delta V}\right)\Lambda_1\rho+\left(1-\frac{P_1^P}{\beta_L\delta V}\right)\left(\Lambda_1+\Lambda_0\right)(1-\rho).$$

Under sufficient conditions (the necessary condition is difficult to interpret) that $\frac{\Lambda_1}{\Lambda_0}<\frac{1}{2}$, and $\frac{\beta_H}{\beta_L}>\frac{2\rho^2}{(1-2\rho)^2}$, we have

$$\lim_{\frac{\delta}{\alpha}\to 0}\frac{\partial Q^P}{\partial\rho}=\frac{\left\{\begin{array}{c}\beta_H^2(\Lambda_0-2\Lambda_1)\left(\Lambda_1+\Lambda_0\right)^2(1-\rho)^2+\Lambda_1^3\beta_L\left[\beta_H(1-2\rho)^2-2\beta_L\rho^2\right]\\ +\beta_H\beta_L\Lambda_1\Lambda_0^2(1-\rho^2)+\beta_H\beta_L\Lambda_1^2\Lambda_0\left(2-4\rho+3\rho^2\right)\end{array}\right\}}{2\left[\beta_H\left(\Lambda_1+\Lambda_0\right)(1-\rho)+\beta_L\Lambda_1\rho\right]^2}>0.$$

In equilibrium-D, the aggregate product adoption is

$$Q^D=\Lambda_0+\Lambda_1\left[\left(1-\frac{P_1^D}{\beta_H\delta V}\right)\rho+\left(1-\frac{P_1^D}{\beta_L\delta V}\right)(1-\rho)\right].$$

It is straightforward to check that

$$\frac{\partial Q^D}{\partial\rho}=\frac{\beta_H\beta_L^2\hat{\beta}^2(2\hat{\beta}-1)\left(\beta_H-\beta_L\right)\delta\Lambda_0\Lambda_1(\beta_L\delta\Lambda_0+\alpha\Lambda_1)}{\left\{\beta_H\beta_L^2(2\hat{\beta}-1)\delta\Lambda_0+2\alpha\hat{\beta}^2\Lambda_1\left[\beta_H(1-\rho)+\beta_L\rho\right]\right\}^2},$$

which implies that $\hat{\beta}>\frac{1}{2}\Leftrightarrow\frac{\partial Q^D}{\partial\rho}>0$.

Finally, in equilibrium-N, the aggregate product adoption is

$$Q^N=\left[\left(1-\frac{P_1^N}{\beta_H\delta V}\right)\rho+\left(1-\frac{P_1^N}{\beta_L\delta V}\right)(1-\rho)\right]\left(\Lambda_0+\Lambda_1\right)=\frac{\Lambda_0+\Lambda_1}{2},$$

and $\frac{\partial Q^N}{\partial \rho} = 0$. □

Proof of Proposition 19. The proof follows the same procedure as Proposition 13 and we omit the algebra. The results for adoption quantities are easily checked by observing that

$$Q^D = \Lambda_0 + \Lambda_1 \left[\frac{1}{2} - \frac{\beta(2-\beta)\delta\Lambda_0}{4\alpha\Lambda_1 + 2\beta^2\delta\Lambda_0}\right] < \Lambda_0 + \frac{\Lambda_1}{2} = \tilde{Q}^D$$

$$Q^P = \gamma\Lambda_0 + [(1-\gamma)\Lambda_0 + \Lambda_1]\left[\frac{1}{2} - \frac{\gamma\delta\Lambda_0}{4\alpha\left[(1-\gamma)\Lambda_0 + \Lambda_1\right] - 2\gamma\left(1-2\beta\right)\delta\Lambda_0}\right] < \tilde{Q}^P,$$

(as long as $P_1^P > 0$), and $Q^N = \frac{\Lambda_0+\Lambda_1}{2} = \tilde{Q}^N$. □

Proof of Proposition 20. Consider the case when the donor subsidizes in the spot-period. For the separating equilibrium, the seller maximizes the spot-period revenue by choosing $\tilde{P}_1^P = \frac{s+\beta\delta V}{2}$. The binding incentive compatibility constraint sets $\tilde{P}_0^P = \frac{\beta\delta^2 V}{2} - \int_{\frac{\tilde{P}_1^P - s}{\beta\delta V}}^{1} \left[\beta\delta^2\theta V - \beta\delta\left(\tilde{P}_1^P - s\right)\right] d\theta$. The aggregate product adoption quantity becomes

$$\tilde{Q}^P(s) = \gamma\Lambda + [(1-\gamma)\Lambda_0 + \Lambda_1]\left(1 - \frac{\tilde{P}_1^P - s}{\beta\delta V}\right)$$

The marginal impact of unit subsidy is

$$\frac{\partial\tilde{Q}^P(s)}{\partial s} = \frac{[(1-\gamma)\Lambda_0 + \Lambda_1]}{2\beta\delta V} = \frac{\alpha\left[(1-\gamma)\Lambda_0 + \Lambda_1\right] - \frac{\gamma(1-2\beta)\delta\Lambda_0}{2}}{\alpha\left[(1-\gamma)\Lambda_0 + \Lambda_1\right]} \cdot \frac{\partial Q^P(s)}{\partial s}.$$

Thus, $\frac{\partial\tilde{Q}^P(s)}{\partial s} > \frac{\partial Q^P(s)}{\partial s}$ if and only if $\beta > \frac{1}{2}$.

For equilibrium-D, $\tilde{P}_1^D = \frac{s+\beta\delta V}{2}$ $\tilde{P}_0^D = \frac{\beta\delta^2 V}{2} - \int_{\frac{\tilde{P}_1^D - s}{\delta V}}^{1} \left[\beta\delta^2\theta V - \beta\delta\left(\tilde{P}_1^D - s\right)\right] d\theta$. The adoption quantity becomes $\tilde{Q}^D(s) = \Lambda_0 + \Lambda_1\left(\frac{1}{2} + \frac{s}{2\beta\delta V}\right)$, and

$$\frac{\partial\tilde{Q}^D(s)}{\partial s} = \frac{\Lambda_1}{2\beta\delta V} = \frac{2\alpha\Lambda_1 + \beta^2\delta\Lambda_0}{2\alpha\Lambda_1} \cdot \frac{\partial Q^D(s)}{\partial s} > \frac{\partial Q^D(s)}{\partial s}.$$

While in equilibrium-N, $\tilde{Q}^N(s) = Q^N(s)$, and $\frac{\partial\tilde{Q}^N(s)}{\partial s} = \frac{\partial Q^N(s)}{\partial s} = \frac{\Lambda_0+\Lambda_1}{4\beta\delta V}$. □

Appendix C

Appendix for Chapter 4

In this appendix, we provide the detailed proofs of the main results in Chapter 4.

Proof of proposition 21. The first statement follows because the objective function is linear to the expected delay for all priority classes after plugging in the pricing solutions. From classical queueing scheduling theory, we know that the optimal scheduling policy is work-conserving, and that the $c\mu$ rule applies. In the first queue for instance, $\frac{\partial \Pi/\partial W_L^1}{A_L^1} = C_L = \frac{\partial \Pi/\partial W_{Lf}^1}{A_{Lf}^1}$, and $\frac{\partial \Pi/\partial W_H^1}{A_H^1} = C_H + \frac{A_L^1}{A_H^1}(C_H - C_L) > C_L$, $\frac{\partial \Pi/\partial W_{Hf}^1}{A_{Hf}^1} = C_H + \frac{A_{Lf}^1}{A_{Hf}^1}(C_H - C_L) > C_L$. We know that the order of these coefficients is the same as order of queueing priority, which leads to statement 2. □

Proof of proposition 22. The only difference with the server-specific model is that service provider is in charge of routing as an additional control. Now that we no longer have the (ID) constraints, we need the following IC constraints to induce the truth-telling of the customers' flexibility:

$$u_f(i|i) \geq u_m(i|i), \forall i \in T, \forall m \in M. \tag{C.1}$$

This implies $u_1(i|i)r_i + u_2(i|i)(1 - r_i) \geq u_m(i|i), \forall i \in T, \forall m \in M$, which is equivalent to $u_1(i|i) = u_2(i|i) = u_f(i|i)$, $\forall i \in T$, for all the interior solutions of routing probabilities. In other words, we equivalently have the same (ID) constraints, and thus this new mechanism is equivalent to the server-specific mechanism. □

Proof of lemma 4. The IC constraints can be equivalently rewritten as:

$$(C_H - C_L)q_L^m W_L^m \geq u_m(L|L) - u_m(H|H) \geq (C_H - C_L)q_H^m W_H^m, \forall m \in M. \tag{C.2}$$

Since whenever $q_H^m = 1$, we have $u_m(L|L) \geq u_m(H|H)$ and thus $q_L^m = 1$, $\forall m \in M$, which means that the service provider will not serve only impatient customers. Next, we need to prove that IC constraints for the flexible customers are redundant as could be shown below in an equivalent form:

$$(C_H - C_L)[q_L^1 W_L^1 r_L + q_L^2 W_L^2 (1 - r_L)] \geq u_f(L|L) - u_f(H|H) \geq (C_H - C_L)[q_H^1 W_H^1 r_H + q_H^2 W_H^2 (1 - r_H)]. \tag{C.3}$$

If we multiply both sides of the IC constraints for the dedicated customers with r_L when $m = 1$ and with $(1 - r_L)$ for $m = 2$, and summing these up should yield:

$$(C_H - C_L)(q_L^1 W_L^1 r_L + q_L^2 W_L^2 (1 - r_L)) \geq (u_1(L|L) - u_1(H|H))r_L + (u_2(L|L) - u_2(H|H))(1 - r_L). \tag{C.4}$$

Due to the ID constraints $u_1(i|i) = u_2(i|i) = u_f(i|i)$, $\forall i \in T$, we have:

$$(C_H - C_L)(q_L^1 W_L^1 r_L + q_L^2 W_L^2 (1 - r_L)) \geq u_f(L|L) - u_f(H|H). \tag{C.5}$$

Similar argument for the second half of IC constraints would yield:

$$u_f(L|L) - u_f(H|H) \geq (C_H - C_L)(q_H^1 W_H^1 r_H + q_H^2 W_H^2 (1 - r_H)). \tag{C.6}$$

By combining the above two inequalities we conclude that the IC constraints for the flexible customers are implied by IC constraints for the dedicated customers. □

Proof of lemma 5. From the IC constraints for the L-type customers we know that $P_H^m - P_L^m \geq C_L W_H^m + C_L W_L^m$, $\forall m \in \{1, 2\}$. Plugging this into $u_m(L|L) - u_m(H|H) = C_L W_L^m + C_H W_H^m + P_H^m - P_L^m$, we can conclude that the utility surplus from the L-type customers is strictly greater than that from the H-type customers:

$$u_m(L|L) - u_m(H|H) = (C_H - C_L) W_H^m > 0, \forall m \in \{1, 2\}. \tag{C.7}$$

Next, we prove by contradiction for the fact that the IR constraints for the H-type customers are binding. Suppose this is not true for the optimal solution triples (P_i^m, W_i^m, r_i), and assume that $\forall \delta > 0$ such that $V - C_H W_H^1 - P_H^1 = \delta > 0$. Then $V - C_L W_L^1 - P_L^1 > \delta > 0$, by the ordering of the quantities of the utility surplus. From the ID constraints we know that this is also true for the second queue: $V - C_H W_H^2 - P_H^2 = V - C_H W_H^1 - P_H^1 = \delta > 0$ and $V - C_L W_L^2 - P_L^2 > \delta > 0$. Now we add small variations to prices, and denote $\widetilde{P}_i^m = P_i^m + \epsilon, \forall i \in \{H, L\}$, $\forall m \in \{1, 2\}$, where $\epsilon > 0$. It could be checked that this new pricing strategy is feasible since we increase both sides of all IC constraints by ϵ without making any adjustments to other variables. This constructed pricing strategy would produce additional revenue for the service provider of $(\sum_{i,m} A_i^m)\epsilon$, contradictory to the assumption that the original pricing strategy is optimal.

Now we prove that the IR constraints for H-type customers are binding, and thus the ID constraints for the H-type customers are binding trivially. We are left with only two IC constraints and one ID constraints for the L-type customers. For given arrival rates and scheduling policies, the optimal pricing strategies would live on a simplex where two constraints out of the remaining three being active. Therefore, at least one of the IC constraints should be binding. □

Proof of Proposition 23. We start by solving the first subproblem in three steps.

Step one. We first characterize all feasible solutions. Given traffic assignment profile $r = (r_H, r_L)$, the achievable regions $R^m(r)$ on the two-dimensional plain of (W_L^m, W_H^m),

$\forall m \in M$, are defined by the resource constraints, where the expected delay for each customer segment is function of traffic assignment profile r:

$$R^m(r) = \left\{ (W_L^m, W_H^m) | \sum_{i \in S} A_i^m W_i^m(r) \geq \frac{\sum_{i \in S} A_i^m}{1 - \sum_{i \in S} A_i^m}, \forall m \in M, \forall S \subset T \right\}, \forall m \in M. \tag{C.8}$$

Among these resource constraints, the *conservation law* for all customers combined requires:

$$A_H^m W_H^m(r) + A_L^m W_L^m(r) \geq \frac{A_H^m + A_L^m}{1 - A_H^m - A_L^m}, \forall m \in M. \tag{C.9}$$

If this work conservation is binding, the delay profile (W_L^m, W_H^m) is conditionally *Pareto efficient*, meaning that there is no waste of resources and no server could be idle if there are customers awaiting to be served. Geometrically as depicted in Figure C.1, the straight line represented by $\{(W_L^m, W_H^m) | \frac{A_H^m}{A_H^m + A_L^m} W_H^m + \frac{A_L^m}{A_H^m + A_L^m} W_L^m = \frac{1}{1 - A_H^m - A_L^m}\}$, $\forall m \in M$ is defined in Yahalom (2006) as *efficient frontier* (EF).

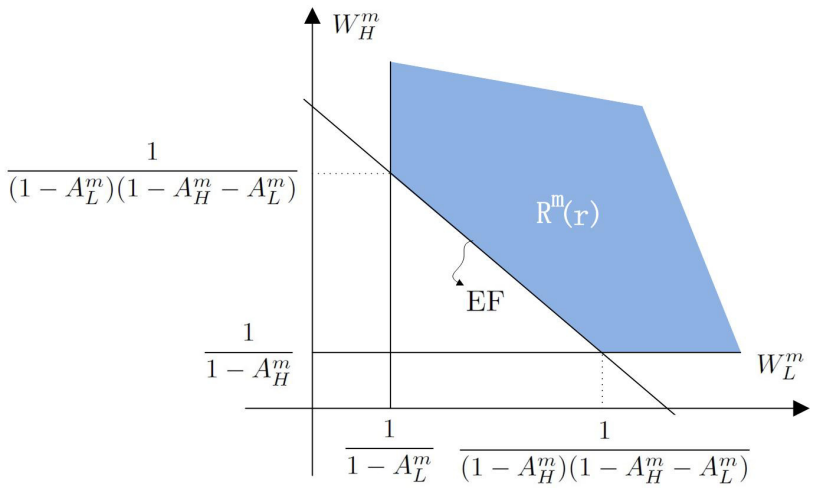

Figure C.1: Achievable region on the two-dimensional plain of delay profile.

If we relax the scheduling constraints "$W_H^2 \leq W_H^1 \leq W_L^2, W_H^1 \leq W_L^1$" implied by incentive compatibility constraints, the $c\mu$ rule implies that $(\frac{1}{(1-A_H^m)(1-A_H^m-A_L^m)}, \frac{1}{1-A_H^m})$, $\forall m \in M$ is the unique optimal solution for the first subproblem. This delay profile corresponds to a scheduling policy that the H-type customers should be given *absolute preemptive priority*.

Step two. We show how the optimal solutions would deviate from $c\mu$ rule, if we take into consideration scheduling constraints $W_H^2 \leq W_H^1 \leq W_L^2$, and $W_H^1 \leq W_L^1$. First by applying $W_H^2 \leq W_H^1 \leq W_L^2$, we partition the traffic rates to both queues into three sets:

$$U_{LH-LH}^1 = \left\{ (A_H^1, A_L^1, A_H^2, A_L^2) : W_H^1 \geq \frac{1}{(1 - A_H^2)(1 - A_H^2 - A_L^2)} \right\}, \tag{C.10}$$

$$U^0_{LH-LH} = \left\{ (A^1_H, A^1_L, A^2_H, A^2_L) : \frac{1}{1-A^2_H} \le W^1_H \le \frac{1}{(1-A^2_H)(1-A^2_H-A^2_L)} \right\}, \tag{C.11}$$

$$U^2_{LH-LH} = \left\{ (A^1_H, A^1_L, A^2_H, A^2_L) : W^1_H \le \frac{1}{1-A^2_H} \right\}. \tag{C.12}$$

For the first subproblem, given the expected delay of the H-type customers in the first queue, the optimal expected delay in the second queue is obtained as follows:

1. When $(A^1_H, A^1_L, A^2_H, A^2_L) \in U^1_{LH-LH}$, the H-type customers in the second queue are given absolute preemptive priority, while the L-type customers in the second queue are given inserted strategic delays, i.e., $W^2_H = \frac{1}{1-A^2_H}$, and $W^2_L = W^1_H > \frac{1}{(1-A^2_H)(1-A^2_H-A^2_L)}$.

2. When $(A^1_H, A^1_L, A^2_H, A^2_L) \in U^0_{LH-LH}$, the H-type customers in the second queue are given absolute preemptive priority, and the scheduling policies would be work-conserving, i.e., $W^2_H = \frac{1}{1-A^2_H}$, and $W^2_L = W^1_H = \frac{1}{(1-A^2_H)(1-A^2_H-A^2_L)}$.

3. When $(A^1_H, A^1_L, A^2_H, A^2_L) \in U^0_{LH-LH}$, there is no feasible scheduling policy for the second queue.

These three observations follow from Figure C.2, where achievable region for the second queue is further restricted by $W^2_H \le W^1_H \le W^2_L$ implied by IC constraints.

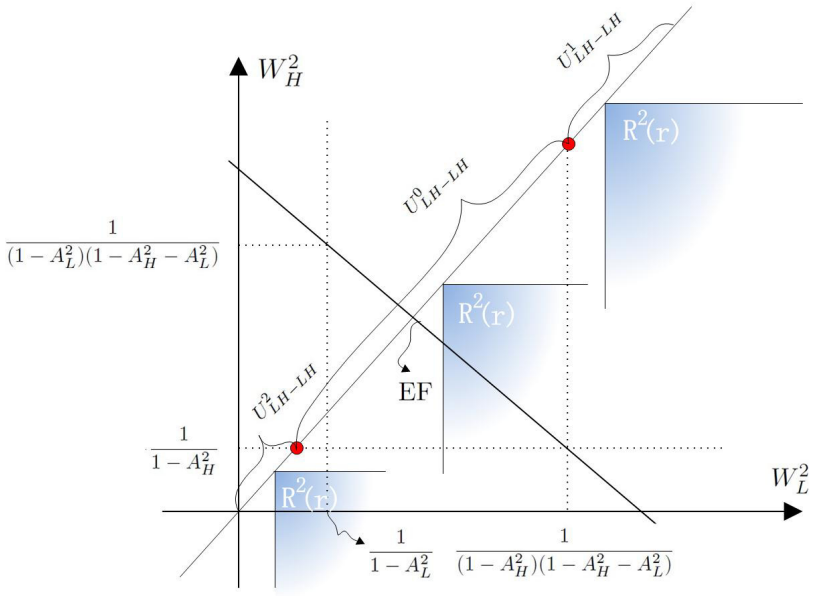

Figure C.2: Achievable region for the second queue restricted by IC.

Step three. Now we study the expected delay profile of the first queue in both feasible scenarios U^1_{LH-LH} and U^0_{LH-LH}. First we discuss the scenario when $(A^1_H, A^1_L, A^2_H, A^2_L) \in U^1_{LH-LH}$ by further elaboration of three cases:

1. If $\frac{1}{1-A_H^1} \leq \frac{1}{(1-A_H^2)}$, we have no feasible solutions.

2. If $\frac{1}{(1-A_H^2)} \leq \frac{1}{1-A_H^1} \leq \frac{1}{(1-A_H^2)(1-A_H^2-A_L^2)}$, we notice that the achievable region is restricted by additional constraints $W_H^1 \geq \frac{1}{(1-A_H^2)(1-A_H^2-A_L^2)}$ and it is binding for the same argument by which we used to show that $c\mu$ rule holds. Furthermore, we can decrease W_H^1 until either $W_L^1 = \frac{1}{(1-A_H^1)(1-A_H^1-A_L^1)}$ or $W_H^1 \geq \frac{1}{(1-A_H^2)(1-A_H^2-A_L^2)}$ is no longer satisfied. In other words, we have $W_H^1 = W_L^2 = \frac{1}{(1-A_H^2)(1-A_H^2-A_L^2)}$, $W_H^2 = \frac{1}{(1-A_H^2)}$, and $W_L^1 = \frac{1}{(1-A_H^1)(1-A_H^1-A_L^1)}$.

3. If $\frac{1}{(1-A_H^2)(1-A_H^2-A_L^2)} \leq \frac{1}{1-A_H^1}$, the additional constraints imposed by IC constraints are no more stringent than RE constraints on the expected delay profile of the first queue. We conclude in this case that $W_H^1 = W_L^2 = \frac{1}{1-A_H^1}$, $W_H^2 = \frac{1}{1-A_H^2}$, and $W_L^1 = \frac{1}{(1-A_H^1)(1-A_H^1-A_L^1)}$.

When $(A_H^1, A_L^1, A_H^2, A_L^2) \in U_{LH-LH}^0$, we discuss the results in the following three cases:

1. If $\frac{1}{1-A_H^1} \leq \frac{1}{(1-A_H^2)}$, the achievable region of the expected delay profile for the first queue is restricted by $W_H^1 \geq \frac{1}{1-A_H^2}$ which is biding, yielding $W_H^1 = W_H^2 = \frac{1}{1-A_H^2}$ and $W_L^2 = \frac{1}{(1-A_H^2)(1-A_H^2-A_L^2)}$. However, W_L^1 depends on the relative value of $\frac{1}{1-A_H^2}$ and $\frac{1}{1-A_H^1-A_L^1}$, which is shown as in Figure C.3.

2. If $\frac{1}{(1-A_H^2)} \leq \frac{1}{1-A_H^1} \leq \frac{1}{(1-A_H^2)(1-A_H^2-A_L^2)}$, then all scheduling constraints implied by IC are redundant.

3. If $\frac{1}{(1-A_H^2)(1-A_H^2-A_L^2)} \leq \frac{1}{1-A_H^1}$, we have no feasible solution.

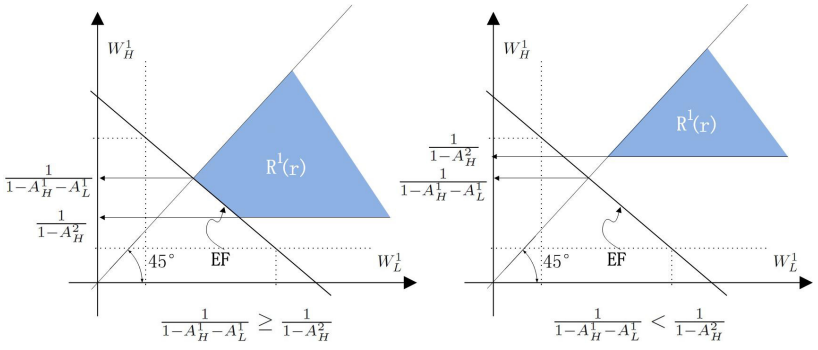

Figure C.3: Achievable region for the first queue restricted in U_{LH-LH}^0, case 1.

To wrap up this discussion, we summarized the results in Table C.1. Symmetrically we derive the optimal scheduling policies in the second subproblem via similar analysis, and the results are shown in Table C.2.

Table C.1: Conditionally optimal scheduling policy for the first subproblem.

Traffic Regime	Expected Steady State Delay	Scheduling Policies
$\frac{1}{1-A_H^1} < \frac{1}{1-A_H^2}$	$W_H^1 = W_H^2 = W_L^1 = \frac{1}{1-A_H^2}$	Strategic delay in queue 1
$\frac{1}{1-A_H^1-A_L^1} < \frac{1}{1-A_H^2}$	$W_L^2 = \frac{1}{(1-A_H^2)(1-A_H^2-A_L^2)}$	Absolute preemptive priority in queue 2
$\frac{1}{1-A_H^1} < \frac{1}{1-A_H^2}$	$W_H^1 = W_H^2 = \frac{1}{1-A_H^2}, W_L^2 = \frac{1}{(1-A_H^2)(1-A_H^2-A_L^2)}$	Randomized preemptive priority in queue 1
$\frac{1}{1-A_H^1-A_L^1} \geq \frac{1}{1-A_H^2}$	$W_L^1 = \frac{1}{A_L^1}(\frac{A_H^1+A_L^1}{1-A_L^1-A_H^1} - \frac{A_H^1}{1-A_H^2})$	Absolute preemptive priority in queue 2
$\frac{1}{1-A_H^1} \geq \frac{1}{1-A_H^2}$	$W_H^1 = \frac{1}{1-A_H^1}, W_L^1 = \frac{1}{(1-A_H^1)(1-A_H^1-A_L^1)}$	Absolute preemptive priority in queue 1
$\frac{1}{1-A_H^1} \leq \frac{1}{(1-A_H^2)(1-A_H^2-A_L^2)}$	$W_H^2 = \frac{1}{1-A_H^2}, W_L^2 = \frac{1}{(1-A_H^2)(1-A_H^2-A_L^2)}$	Absolute preemptive priority in queue 2
$\frac{1}{1-A_H^1} > \frac{1}{(1-A_H^2)(1-A_H^2-A_L^2)}$	$W_L^2 = W_H^1 = \frac{1}{1-A_H^1}, W_H^2 = \frac{1}{1-A_H^2}$	Absolute preemptive priority in queue 1
	$W_L^1 = \frac{1}{(1-A_H^1)(1-A_H^1-A_L^1)}$	Strategic delay in queue 2

Table C.2: Conditionally optimal scheduling policy for the second subproblem.

Traffic Regime	Expected Steady State Delay	Scheduling Policies
$\frac{1}{1-A_H^2} < \frac{1}{1-A_H^1}$	$W_H^2 = W_H^1 = W_L^2 = \frac{1}{1-A_H^1}$	Strategic delay in queue 2
$\frac{1}{1-A_H^2-A_L^2} < \frac{1}{1-A_H^1}$	$W_L^1 = \frac{1}{(1-A_H^1)(1-A_H^1-A_L^1)}$	Absolute preemptive priority in queue 1
$\frac{1}{1-A_H^2} < \frac{1}{1-A_H^1}$	$W_H^2 = W_H^1 = \frac{1}{1-A_H^1}, W_L^1 = \frac{1}{(1-A_H^1)(1-A_H^1-A_L^1)}$	Randomized preemptive priority in queue 2
$\frac{1}{1-A_H^2-A_L^2} \geq \frac{1}{1-A_H^1}$	$W_L^2 = \frac{1}{A_L^2}(\frac{A_H^2+A_L^2}{1-A_L^2-A_H^2} - \frac{A_H^2}{1-A_H^1})$	Absolute preemptive priority in queue 1
$\frac{1}{1-A_H^2} \geq \frac{1}{1-A_H^1}$	$W_H^2 = \frac{1}{1-A_H^2}, W_L^2 = \frac{1}{(1-A_H^2)(1-A_H^2-A_L^2)}$	Absolute preemptive priority in queue 2
$\frac{1}{1-A_H^2} \leq \frac{1}{(1-A_H^1)(1-A_H^1-A_L^1)}$	$W_H^1 = \frac{1}{1-A_H^1}, W_L^1 = \frac{1}{(1-A_H^1)(1-A_H^1-A_L^1)}$	Absolute preemptive priority in queue 1
$\frac{1}{1-A_H^2} > \frac{1}{(1-A_H^1)(1-A_H^1-A_L^1)}$	$W_L^1 = W_H^2 = \frac{1}{1-A_H^2}, W_H^1 = \frac{1}{1-A_H^1}$	Absolute preemptive priority in queue 2
	$W_L^2 = \frac{1}{(1-A_H^2)(1-A_H^2-A_L^2)}$	Strategic delay in queue 1

Combine the solutions in the two subproblems by comparing the maximized revenue, we obtain the results summarized in Table 4.4. □

Proof of proposition 24: For the first statement, consider case 1 in Table 4.4. Plugging in the pricing and the expected delay solutions, we can write down the total revenue in its

closed form:

$$\Pi_{case1} = (\lambda_H^1 + \lambda_H^2 + \lambda_H^f + \lambda_L^1 + \lambda_L^2 + \lambda_L^f)V + C_H \left(2 - \frac{1 + \lambda_L^1 + \lambda_L^2 + \lambda_L^f}{1 - \lambda_H^2 - \lambda_H^f(1 - r_H)} - \frac{1}{1 - \lambda_H^1 - \lambda_H^f r_H}\right) + \frac{C_L(\lambda_L^2 + \lambda_L^f(1 - r_L))(\lambda_H^2 + \lambda_L^2 + \lambda_H^f(1 - r_H) + \lambda_L^f(1 - r_L))}{1 - \lambda_H^2 - \lambda_H^f(1 - r_H)}. \tag{C.13}$$

Consider:

$$\begin{aligned}\frac{\partial \Pi_{case1}}{\partial r_L} &= \frac{C_L\lambda_L^f}{(1 - \lambda_H^2 - \lambda_H^f(1 - r_H) - \lambda_L^2 - \lambda_L^f(1 - r_L))^2} - \frac{C_L\lambda_L^f}{1 - \lambda_H^2 - \lambda_H^f(1 - r_H)} \\ &\geq \frac{C_L\lambda_L^f}{(1 - \lambda_H^2 - \lambda_H^f(1 - r_H) - \lambda_L^2 - \lambda_L^f(1 - r_L))} - \frac{C_L\lambda_L^f}{1 - \lambda_H^2 - \lambda_H^f(1 - r_H)} \geq 0.\end{aligned} \tag{C.14}$$

Note that the denominators in the above equations are all positive due to the stability constraints. The inequality follows by directly comparing the denominators. Since $\frac{\partial \Pi}{\partial r_L} \geq 0$ holds for all $r_L \in (0, 1)$, thus we have $r_L \to 1$.

On the other hand, consider:

$$\begin{aligned}\frac{\partial \Pi_{case1}}{\partial r_H} &= \frac{C_H\lambda_H^f(1 + \lambda_L^1 + \lambda_L^2 + \lambda_L^f)}{(1 - \lambda_H^2 - \lambda_H^f(1 - r_H))^2} - \frac{C_H\lambda_H^f}{(1 - \lambda_H^1 - \lambda_H^f)^2} \\ &\quad + \frac{C_L\lambda_H^f}{(1 - \lambda_H^2 - \lambda_L^2 - \lambda_H^f(1 - r_H) - \lambda_L^f(1 - r_L))^2} - \frac{C_L\lambda_H^f(1 + \lambda_L^2 + \lambda_L^f(1 - r_L))}{(1 - \lambda_H^2 - \lambda_H^f(1 - r_H))^2}.\end{aligned} \tag{C.15}$$

Plug in $r_L \to 1$, and we have:

$$\begin{aligned}\frac{\partial \Pi_{case1}}{\partial r_H} &= \frac{C_H\lambda_H^f(1 + \lambda_L^1 + \lambda_L^2 + \lambda_L^f)}{(1 - \lambda_H^2 - \lambda_H^f(1 - r_H))^2} - \frac{C_H\lambda_H^f}{(1 - \lambda_H^1 - \lambda_H^f)^2} \\ &\quad + \frac{C_L\lambda_H^f}{(1 - \lambda_H^2 - \lambda_L^2 - \lambda_H^f(1 - r_H))^2} - \frac{C_L\lambda_H^f(1 + \lambda_L^2)}{(1 - \lambda_H^2 - \lambda_H^f(1 - r_H))^2} \\ &\geq \left[\frac{C_H\lambda_H^f}{(1 - \lambda_H^2 - \lambda_H^f(1 - r_H))^2} - \frac{C_H\lambda_H^f}{(1 - \lambda_H^1 - \lambda_H^f)^2}\right] \\ &\quad + \left[\frac{C_L\lambda_H^f}{(1 - \lambda_H^2 - \lambda_L^2 - \lambda_H^f(1 - r_H))^2} - \frac{C_L\lambda_H^f}{(1 - \lambda_H^2 - \lambda_H^f(1 - r_H))^2}\right] \\ &\quad + \left[\frac{C_H\lambda_H^f(\lambda_L^1 + \lambda_L^f) + \lambda_H^f\lambda_L^2(C_H - C_L)}{(1 - \lambda_H^2 - \lambda_H^f(1 - r_H))^2}\right] \geq 0.\end{aligned} \tag{C.16}$$

where in the last inequality, by the assumption that $\lambda_H^2 > \lambda_H^1 + \lambda_H^f + \lambda_L^1 + \lambda_L^f + (\lambda_H^1 + \lambda_H^f)(1 - \lambda_H^1 - \lambda_H^f - \lambda_L^1 - \lambda_L^f)$, we have $\lambda_H^2 \geq \lambda_H^1 + \lambda_H^f r_H$ and the first item is nonnegative. The second item is nonnegative simply by comparing denominators. Now that we have shown that $\frac{\partial \Pi}{\partial r_H} \geq 0$ holds for all $r_H \in (0,1)$, it must be that $r_L \to 1$.

The argument for the second case is similar. First we can plug in the expected delay and pricing solutions and write down the service provider's revenue as follows:

$$\begin{aligned}\Pi_{case2} = (\lambda_H^1 + \lambda_H^2 + \lambda_H^f + \lambda_L^1 + \lambda_L^2 + \lambda_L^f)V - \frac{C_H(\lambda_H^1 + \lambda_H^2 + \lambda_H^f + \lambda_L^1 + \lambda_L^2 + \lambda_L^f)}{1 - \lambda_H^2 - \lambda_H^f(1 - r_H)} \\ - \frac{C_L(\lambda_L^2 + \lambda_L^f - \lambda_L^f r_L)(\lambda_H^2 + \lambda_H^f + \lambda_L^2 + \lambda_L^f - \lambda_H^f r_H - \lambda_L^f r_L)}{(1 - \lambda_H^2 - \lambda_H^f(1 - r_H))(1 + \lambda_H^2 + \lambda_H^f + \lambda_L^2 + \lambda_L^f - \lambda_H^f r_H - \lambda_L^f r_L)}.\end{aligned} \tag{C.17}$$

Take derivative:

$$\frac{\partial \Pi_{case2}}{\partial r_L} = \frac{C_L \lambda_L^f}{(1 - \lambda_H^2 - \lambda_H^f(1 - r_H) - \lambda_L^2 - \lambda_L^f(1 - r_L))^2} - \frac{C_L \lambda_L^f}{1 - \lambda_H^2 - \lambda_H^f(1 - r_H)} \geq 0. \tag{C.18}$$

Note that the $\frac{\partial \Pi}{\partial r_L}$ in this case is exactly the same with the first case and thus the optimal routing probability $r_L \to 1$, by using exactly the same argument. To see $r_H \to 1$, we need to check $\frac{\partial \Pi}{\partial r_H} \geq 0$ for $\forall r_H \in (0,1)$:

$$\begin{aligned}\frac{\partial \Pi_{case2}}{\partial r_H} = \frac{(C_H - C_L)(1 + \lambda_L^2 + \lambda_L^f)\lambda_H^f + C_H \lambda_H^f(\lambda_H^1 + \lambda_H^2 + \lambda_H^f) + C_L \lambda_H^f \lambda_L^f r_L}{(1 - \lambda_H^2 - \lambda_H^f(1 - r_H))^2} \\ + \frac{C_L \lambda_H^f}{(1 - \lambda_H^2 - \lambda_H^f(1 - r_H) - \lambda_L^2 - \lambda_L^f(1 - r_L))^2} \geq 0.\end{aligned} \tag{C.19}$$

We can do the same for case 3. Due to limited space, we omit the closed-form solution of Π_{case3}. The derivatives are as follows:

$$\frac{\partial \Pi_{case3}}{\partial r_L} = \frac{C_L \lambda_L^f}{(1 - \lambda_H^2 - \lambda_H^f(1 - r_H) - \lambda_L^2 - \lambda_L^f(1 - r_L))^2} - \frac{C_L \lambda_L^f}{(1 - \lambda_H^1 - \lambda_L^1 - \lambda_H^f r_H - \lambda_L^f r_L)^2}. \tag{C.20}$$

$$\frac{\partial \Pi_{case3}}{\partial r_H} = \frac{\lambda_H^f(C_H - C_L)(1 + \lambda_L^1 + \lambda_L^2 + \lambda_L^f)}{(1 - \lambda_H^2 - \lambda_H^f(1 - r_H))^2} - \frac{\lambda_H^f(C_H - C_L)}{(1 - \lambda_H^1 - \lambda_H^f r_H)^2} + \frac{\lambda_H^f}{\lambda_L^f} \frac{\partial \Pi_{case3}}{\partial r_L}. \tag{C.21}$$

By the first order condition we immediately have $\frac{\partial \Pi_{case3}}{\partial r_L} = 0$, thus $(1 - \lambda_H^2 - \lambda_H^f(1 - r_H) - \lambda_L^2 - \lambda_L^f(1 - r_L)) = (1 - \lambda_H^1 - \lambda_L^1 - \lambda_H^f r_H - \lambda_L^f r_L)$, yielding to the optimal routing probability $r_L = \frac{1}{2} + \frac{\lambda_H^2 + \lambda_L^2 - \lambda_H^1 - \lambda_L^1 - \lambda_H^f}{2\lambda_L^f}$. The sufficient conditions ensure that $r_L \in (0,1)$. Therefore:

$$\begin{aligned}\frac{\partial \Pi_{case3}}{\partial r_H} &= \frac{\lambda_H^f(C_H - C_L)(1+\lambda_L^1+\lambda_L^2+\lambda_L^f)}{(1-\lambda_H^2-\lambda_H^f(1-r_H))^2} - \frac{\lambda_H^f(C_H-C_L)}{(1-\lambda_H^1-\lambda_H^f r_H)^2} \\ &\geq \frac{\lambda_H^f(C_H-C_L)(\lambda_L^1+\lambda_L^2+\lambda_L^f)}{(1-\lambda_H^2-\lambda_H^f(1-r_H))^2} \geq 0.\end{aligned} \tag{C.22}$$

where the last inequality is due to $A_H^2 \geq A_H^1$ or $\lambda_H^f < \lambda_H^2 - \lambda_H^1$. This means at optimality $r_H \to 1$.

We show the last statement (the special case) by contradiction. Suppose $A_H^1 \neq A_H^2$, and we let $A_H^1 < A_H^2$ without loss of generality. Then this would be reduced to case 3 and $\frac{\partial \Pi_{case3}}{\partial r_H} > 0$. Plugging $r_H \to 1$ into $A_H^1 < A_H^2$, we have $\lambda_H^f < \lambda_2^f - \lambda_1^f \geq |\lambda_H^2 - \lambda_H^1|$, which is a contradiction to the assumptions. Therefore, for the special case, there must be $A_H^1 = A_H^2$ at optimality and thus $r_H = \frac{1}{2} + \frac{\lambda_H^2 - \lambda_H^1}{2\lambda_H^f}$, $r_L = \frac{1}{2} + \frac{\lambda_L^2 - \lambda_L^1}{2\lambda_L^f}$. □

Analytical Results for L-L and LH-L Cases. Under L-L admission policy, no H-type customers are served.

Proposition 33 *In the L-L case, given scheduling policy, the conditionally optimal pricing schemes are summarized as follows:*

$$P_L^m = V - C_L W_L^m, \forall m \in \{1,2\}. \tag{C.23}$$

The conditionally optimal scheduling policy would always be work-conserving, and the expected steady state delay for L-type customers is $\frac{1}{1-A_L^m}$,$\forall m \in \{1,2\}$.

Proof: Since the service provider does not accept any H-type customers, the remaining IC and IR constraints for L-type arrivals are listed as follows:

$$\begin{aligned}&(\mathbf{IC}): u_m(L|H) \leq 0, \forall m \in M, \\ &(\mathbf{ID}): u_1(L|L) = u_2(L|L), \\ &(\mathbf{IR}): u_m(L|L) \geq 0, \forall m \in M.\end{aligned} \tag{C.24}$$

At optimality, IR constraints would be binding. We prove this by contradiction. Suppose that this is not true for the optimal triples (P_L^m, W_L^m, r_L), and we assume that there exists a small $\delta > 0$ such that $V - C_L W_L^1 - P_L^1 > \delta$. By ID constraint we know that such is also true for the second queue: $V - C_L W_L^2 - P_L^2 = V - C_L W_L^1 - P_L^1 > \delta$. IC constraints disincentivizing H-type from pretending to be L-type could always be satisfied by enforcing: $\delta < (C_H - C_L) W_L^m$, $\forall m \in M$. Now we make small variations to prices, and denote $\widetilde{P}_L^m = P_L^m + \delta$. The constructed pricing strategy would produce additional revenue by $(A_L^1 + A_L^2)\delta > 0$, while all IC, IR and ID constraints are still satisfied, which is a contradiction to the assumption that P_L^m is optimal.

Since now we know IR constraints for L-type are binding, the conditionally optimal pricing is available immediately. The problem is reduced to:

$$\begin{aligned} \text{Maximize } \Pi &= A_L^1 P_L^1 + A_L^2 P_L^2, \\ \text{subject to } P_L^1 &= V - C_L W_L^1, \\ P_L^2 &= V - C_L W_L^2, \\ W_L^m &\geq \frac{1}{1 - A_L^m}, (ST), \forall m \in M. \end{aligned}$$

and immediately we have $W_L^m = \frac{1}{1-A_L^m}, \forall m \in M$. $\square$

The message we receive from this simple scenario is that, if we shut down the channel for serving H-type customers, nothing would prevent us from extracting all revenue from the L-type customers. On the other hand, since admitted customers are homogeneous, the *conservation law* requires that no server idleness is ever allowed.

Proposition 34 *In the LH-L case, given scheduling policy, if conditions* $W_H^1 \leq W_L^2, W_H^1 \leq W_L^1$ *are satisfied, the conditionally optimal prices are given as follows:*

$$P_H^1 = V - C_H W_H^1, \tag{C.25}$$

$$P_L^m = V - (C_H - C_L) W_H^1 - C_L W_L^m, \forall m \in \{1, 2\}. \tag{C.26}$$

and the optimal expected steady state delays in both queues are summarized in Table C.3:

Table C.3: Conditionally optimal scheduling policy for the LH-L scenario.

Traffic Regime	Expected Steady State Delay	Scheduling Policies
$\frac{1}{1-A_H^1} \leq \frac{1}{1-A_L^2}$	$W_H^1 = \frac{1}{1-A_H^1}, W_L^2 = \frac{1}{1-A_L^2}$	Absolute preemptive priority in queue 1
	$W_L^1 = \frac{1}{(1-A_H^1)(1-A_H^1-A_L^1)}$	Work-conserving in queue 2
$\frac{1}{1-A_H^1} > \frac{1}{1-A_L^2}$	$W_H^1 = W_L^2 = \frac{1}{1-A_H^1}$	Absolute preemptive priority in queue 1
	$W_L^1 = \frac{1}{(1-A_H^1)(1-A_H^1-A_L^1)}$	Strategic delay in queue 2

Proof: Step one. To solve for the conditionally optimal pricing schemes, since all customers are admitted into the first queue while no H-type customer is admitted into the second queue, the remaining IC and IR constraints are listed as follows:

$$\begin{aligned} (IC - H1) &: u_1(H|H) \geq u_1(L|H), \\ (IC - L1) &: u_1(L|L) \geq u_1(H|L), \\ (IC - H2) &: u_2(L|H) \leq 0, \\ (ID - H, IR - H) &: u_1(H|H) = 0, \\ (IR - L) &: u_m(L|L) \geq 0, \forall m \in M. \end{aligned} \tag{C.27}$$

From ID-H we have $P_H^1 = V - C_H W_H^1$. Next, we claim that IC-L1 constraint would be binding at optimality, which yields to $P_L^1 = V - (C_H - C_L) W_H^1 - C_L W_L^1$. Similar to

the approach used in *LH-LH* and *L-L* scenarios, we prove this claim by contradiction. Suppose that there exist a small δ such that $u_1(L|L) - u_1(H|L) = \delta > 0$, meaning that $P_L^1 = V-(C_H-C_L)W_H^1-C_LW_L^1-\delta$. From ID-L we have $P_L^2 = V-(C_H-C_L)W_H^1-C_LW_L^2-\delta$, $u_1(L|L) = u_2(L|L) = (C_H - C_L)W_H^1 + \delta > 0$. Now it remains to be checked that IC-H1 and IC-H2 are satisfied. For IC-H1, we have $u_1(H|H)-u_1(L|H) = (C_H-C_L)(W_L^1-W_H^1)-\delta \geq 0$, due to the assumption that $W_L^1 > W_H^1$ and the fact that we can always choose δ to be small enough by continuity. For IC-H2, we have $u_2(L|H) = (C_H - C_L)(W_H^1 - W_L^2) + \delta > 0$, due to the assumption that $W_H^1 \geq W_L^2$. Now that we have constructed a new pricing solution, with which the total revenue would decrease by $(A_L^1 + A_L^2)\delta$. This indicates that if IC-L1 were not binding, the service provider would be worse off using the only feasible alternative pricing scheme.

When IC-L1 constraint is binding, the same reasoning as above would still hold except that we let δ goes to zeros, resulting in the unique feasible pricing scheme that are optimal given expected delay profile.

Step two. Now that we have conditionally optimal pricing solutions, we need to find the conditionally optimal scheduling policies. By plugging the pricing results, the problem could be rewritten as follows:

$$\begin{aligned}
\text{Maximize } \Pi &= A_H^1P_H^1 + A_L^1P_L^1 + A_L^2P_L^2,\\
\text{subject to } P_H^1 &= V - C_HW_H^1,\\
P_L^m &= V - (C_H - C_L)W_H^1 - C_LW_L^m, \forall m \in \{1,2\},\\
W_H^1 &\leq W_L^2, W_H^1 \leq W_L^1,\\
W_i^m &\geq 0, (ST), (RE), \forall m \in M, \forall i \in T.
\end{aligned}$$

Similar to the procedure for scenario *LH-LH*, we claim that $c\mu$ rule holds without scheduling constraints $W_H^1 \leq W_L^2$ and $W_H^1 \leq W_L^1$, resulting in the optimal solution for the expected delay profile (W_L^1, W_H^1, W_L^2) as $(\frac{1}{(1-A_H^1)(1-A_H^1-A_L^1)}, \frac{1}{1-A_H^2}, \frac{1}{1-A_L^2})$. Notice that: $(\frac{\partial\Pi}{\partial W_L^1}, \frac{\partial\Pi}{\partial W_H^1}, \frac{\partial\Pi}{\partial W_L^2})= -(A_L^1C_L,\ A_H^1C_H + (A_L^1 + A_L^2)(C_H - C_L),\ A_L^2C_L) \prec \mathbf{0}$. Thus We need to check $(\frac{\partial\Pi}{\partial W_L^1}, \frac{\partial\Pi}{\partial W_H^1}) \cdot \mathbf{d_m} < 0$, where $\mathbf{d} = (-A_H^1, A_L^1)^T$:

$$\begin{aligned}
\left(\frac{\partial\Pi}{\partial W_L^1}, \frac{\partial\Pi}{\partial W_H^1}\right) \cdot \mathbf{d} &= -(A_L^1C_L, A_H^1C_H + (A_L^1 + A_L^2)(C_H - C_L)) \cdot (-A_H^1, A_L^1)^T,\\
&= -A_L^1(A_H^1 + A_L^1 + A_L^2)(C_H - C_L) < 0.
\end{aligned} \tag{C.28}$$

Next, we take into consideration scheduling constraints $W_H^1 \leq W_L^2$ and $W_H^1 \leq W_L^1$. The restriction of achievable region on the plain of (W_L^1, W_H^1) is bounded from above, and thus there are two possibilities: if $W_L^2 \geq \frac{1}{1-A_H^1}$, we get the same solution induced by $c\mu$ rule; if $W_L^2 < \frac{1}{1-A_H^1}$, there is no feasible solution. The restriction of achievable region for W_L^2 on the other hand, is one dimensional: if $W_H^1 \leq \frac{1}{1-A_L^2}$, we get the same solution $W_L^2 = \frac{1}{1-A_L^2}$; if

$W_H^1 > \frac{1}{1-A_L^2}$, there are inserted delays for the second queue and $W_L^2 = W_H^1$. To summarize the above discussion, we get results in Table C.3. □

Additional Results: Self-Adaptive Routing

Corollary 11 *In the L-L scenario, if* $\lambda_L^f \geq |\lambda_L^2 - \lambda_L^1|$, *optimal routing probability* $r_L = \frac{1}{2} + \frac{\lambda_L^2 - \lambda_L^1}{2\lambda_L^f}$; *if* $\lambda_L^f < \lambda_L^2 - \lambda_L^1$, *equilibrium queue-joining probability* $r_L \to 1$; *otherwise, when* $\lambda_L^f < \lambda_L^1 - \lambda_L^2$, *equilibrium queue-joining* $r_L \to 0$.

Proof: Total revenue:

$$\Pi = (A_L^1 + A_L^2)V + \frac{A_L^1 C_L}{1 - A_L^1} + \frac{A_L^2 C_L}{1 - A_L^2}. \tag{C.29}$$

and the corresponding derivative:

$$\frac{\partial \Pi}{\partial r_L} = \frac{C_L \lambda_L^f}{(1 - \lambda_L^2 - \lambda_L^f(1 - r_L))^2} - \frac{C_L \lambda_L^f}{(1 - \lambda_L^1 - \lambda_L^f r_L)^2}. \tag{C.30}$$

First order condition $\frac{\partial \Pi}{\partial r_L} = 0$ would be enough to show that the equilibrium queue-joining probability $r_L = \frac{1}{2} + \frac{\lambda_L^2 - \lambda_L^1}{2\lambda_L^f}$. If $\lambda_L^f < \lambda_L^2 - \lambda_L^1$, then we have $\frac{\partial \Pi}{\partial r_L} > 0$ for $\forall r_L \in (0, 1)$ and $r_L \to 1$. Similarly we have $r_L \to 0$ otherwise. □

Corollary 12 *In the LH-L scenario, if customer value* V *is high enough, the equilibrium queue-joining probability* $r_H \to 1$. *If* $\lambda_L^f \geq |\lambda_L^2 - \lambda_H^1 - \lambda_L^1 - \lambda_H^f|$, *the equilibrium queue-joining probability* $r_L = \frac{1}{2} + \frac{\lambda_L^2 - \lambda_H^1 - \lambda_L^1 - \lambda_H^f}{2\lambda_L^f}$; *if* $\lambda_L^f < \lambda_L^2 - \lambda_H^1 - \lambda_L^1 - \lambda_H^f$, $r_L = 1$; *if* $\lambda_L^f < \lambda_H^1 + \lambda_L^1 + \lambda_H^f - \lambda_L^2$, $r_L \to 0$.

Proof: For the traffic regime $\frac{1}{1-A_H^1} \leq \frac{1}{1-A_L^2}$, we have:

$$\frac{\partial \Pi}{\partial r_H} = \lambda_H^f V - \frac{(C_H - C_L)(1 + \lambda_L^1 + \lambda_L^2 + \lambda_L^f)}{(1 - \lambda_H^1 - \lambda_H^f r_H)^2} - \frac{\lambda_H^f C_L}{(1 - \lambda_H^1 - \lambda_L^1 - \lambda_H^f r_H - \lambda_L^f r_L)^2}. \tag{C.31}$$

As long as V is large, we have $\frac{\partial \Pi}{\partial r_H} > 0$ and it is optimal to let $r_H \to 1$. On the other hand:

$$\frac{\partial \Pi}{\partial r_L} = \frac{C_L \lambda_L^f}{(1 - \lambda_L^2 - \lambda_L^f(1 - r_L))^2} - \frac{C_L \lambda_L^f}{(1 - \lambda_H^1 - \lambda_L^1 - \lambda_H^f r_H - \lambda_L^f r_L)^2}. \tag{C.32}$$

By the first order condition, we immediately have $r_L = \frac{1}{2} + \frac{\lambda_L^2 - \lambda_H^1 - \lambda_L^1 - \lambda_H^f}{2\lambda_L^f}$ in equilibrium. If $\lambda_L^f < \lambda_L^2 - \lambda_H^1 - \lambda_L^1 - \lambda_H^f$, $\frac{\partial \Pi}{\partial r_L} > 0$ and $r_L \to 1$. Otherwise we would have the equilibrium queue-joining probability $r_L \to 0$. □

The self-adaptive routing profiles (r_L, r_H) in the *L-L* and *LH-L* scenarios can be derived directly from the solution in *LH-LH* scenario, by setting *H*-type arrival rates to be zero

in the corresponding queue. In other words, the routing solutions in L-L or LH-L scenario could be considered as degenerate cases from the solution in LH-LH scenario. In the L-L scenario, since only L-type customers are admitted, the flexible customers should balance the traffic in both queues. Still, it could be that all flexible customers join a single queue and the two queues are still not balanced. In the LH-L scenario, when valuation is high enough, H-type customers would join the first queue where they would be provided service. Next, the flexible L-type customers would play a mixed strategy so that the aggregated traffic to both queues is balanced.

Appendix D

Appendix for Chapter 5

In this appendix, we provide the detailed proofs of the main results in Chapter 5.

Proof of Proposition 26.

Define a correspondence B as:

$$B(\sigma_H^i, \sigma_L^i, \sigma_H^u, \sigma_L^u) = B_H^i(\sigma_L^i, \sigma_H^u, \sigma_L^u) \times B_L^i(\sigma_H^i, \sigma_H^u, \sigma_L^u) \times B_H^u(\sigma_H^i, \sigma_L^i, \sigma_L^u) \times B_L^u(\sigma_H^i, \sigma_L^i, \sigma_H^u), \tag{D.1}$$

where $B_\theta^i(\cdot)$ and $B_\theta^u(\cdot)$ are the best response functions, i.e.,

$$B_\theta^i(\cdot) = \arg\max_{\sigma \in [0,1]} \sigma W_i(\theta, \cdot),$$

$$B_\theta^u(\cdot) = \arg\max_{\sigma \in [0,1]} \sigma W_u(\theta, \cdot). \tag{D.2}$$

The limiting distribution $\pi(n, V_\varphi; \sigma_H^i, \sigma_L^i, \sigma_H^u, \sigma_L^u)$ is continuous in σ_θ^i and σ_θ^u, $\forall \theta \in \{H, L\}$. In addition, we know that the possible states reside in a closed interval, i.e., $n \in \left[0, \left\lfloor \frac{\mu V_H}{C_L} \right\rfloor\right]$. Thus, by Berge's Theorem, the correspondence B is compact, convex-valued, and upper hemi-continuous. Kakutani's fixed point theorem implies that the map B has a fixed point. □

For convenience in exposition, we first present the proof for Proposition 28. Then, we go back and prove Proposition 27.

Proof of Proposition 28. For the informed customers, their utility of joining the queue is $W_i(n, \theta, V_\varphi) = V_\varphi - (n+1)C_\theta/\mu, \forall \theta \in \{H, L\}, \forall \varphi \in \{H, L\}$. Hence, the informed customers would join if $W_i(n, \theta, V_\varphi) \geqslant 0$, i.e., $n \leqslant \left\lfloor \frac{\mu V_\varphi}{C_\theta} \right\rfloor - 1$. For $n > \left\lfloor \frac{\mu V_\varphi}{C_\theta} \right\rfloor - 1$, the informed customers would balk since $V_\varphi - (\left\lfloor \frac{\mu V_\varphi}{C_\theta} \right\rfloor + 1)C_\theta/\mu < V_\varphi - (\frac{\mu V_\varphi}{C_\theta})C_\theta/\mu = 0$. First, we need to check that this strategy profile could indeed be an equilibrium, provided that n_θ^P exist and are unique for $\forall \theta \in \{H, L\}$.

The uninformed customers would join if

$$W_u(n, \theta, \hat{\gamma}_\theta^P) = \hat{\alpha}_\theta^P(n) W_i(n, \theta, V_H) + (1 - \hat{\alpha}_\theta^P(n)) W_i(n, \theta, V_L) \geqslant 0, \tag{D.3}$$

i.e.,

$$n \leqslant \frac{\mu\left[V_H - \left(1-\hat{\gamma}_\theta^P(n)\right)\left(V_H - V_L\right)\right]}{C_\theta} - 1. \tag{D.4}$$

Let n_θ^P be some integers that violate this inequality. In addition:

$$\begin{aligned}\frac{\mu\left[V_H - \left(1-\hat{\gamma}_\theta^P(n)\right)\left(V_H - V_L\right)\right]}{C_\theta} &= \hat{\alpha}_\theta^P(n)\frac{\mu V_H}{C_\theta} + (1-\hat{\alpha}_\theta^P(n))\frac{\mu V_L}{C_\theta} \\ &= \frac{\mu V_H}{C_\theta}\cdot\frac{\pi_0}{\pi_0 + (1-\pi_0)\frac{\hat{\pi}_\theta^P(n,V_L,\widehat{\sigma}_\theta^i,\widehat{\sigma}_\theta^u)}{\hat{\pi}_\theta^P(n,V_H,\widehat{\sigma}_\theta^i,\widehat{\sigma}_\theta^u)}} \\ &\quad + \frac{\mu V_L}{C_\theta}\cdot\frac{(1-\pi_0)}{\pi_0\frac{\hat{\pi}_\theta^P(n,V_H,\widehat{\sigma}_\theta^i,\widehat{\sigma}_\theta^u)}{\hat{\pi}_\theta^P(n,V_L,\widehat{\sigma}_\theta^i,\widehat{\sigma}_\theta^u)} + (1-\pi_0)},\end{aligned}$$

and this would be a convex combination of $\frac{\mu V_H}{C_\theta}$ and $\frac{\mu V_L}{C_\theta}$, i.e., the uninformed customers would join for $\forall n \leqslant n_\theta^P$, for thresholds $\left\lfloor\frac{\mu V_L}{C_H}\right\rfloor \leqslant n_\theta^P \leqslant \left\lfloor\frac{\mu V_H}{C_H}\right\rfloor$. Under such putatively correct Bayesian updating, limiting distribution is determined by the birth-death process:

$$\hat{\pi}_\theta^P(n,V_L,\widehat{\sigma}_\theta^i,\widehat{\sigma}_\theta^u) = \begin{cases} \frac{\left(\frac{\Lambda}{\mu}\right)^n}{1+\sum_{k=1}^{k=\left\lfloor\frac{\mu V_L}{C_\theta}\right\rfloor}\left(\frac{\Lambda}{\mu}\right)^k + \sum_{k=\left\lfloor\frac{\mu V_L}{C_\theta}\right\rfloor+1}^{k=n_\theta^P}(1-\beta)^{k-\left\lfloor\frac{\mu V_L}{C_\theta}\right\rfloor}\left(\frac{\Lambda}{\mu}\right)^k}, n \leqslant \left\lfloor\frac{\mu V_L}{C_\theta}\right\rfloor \\ \frac{(1-\beta)^{n-\left\lfloor\frac{\mu V_L}{C_\theta}\right\rfloor}\left(\frac{\Lambda}{\mu}\right)^n}{1+\sum_{k=1}^{k=\left\lfloor\frac{\mu V_L}{C_\theta}\right\rfloor}\left(\frac{\Lambda}{\mu}\right)^k + \sum_{k=\left\lfloor\frac{\mu V_L}{C_\theta}\right\rfloor+1}^{k=n_\theta^P}(1-\beta)^{k-\left\lfloor\frac{\mu V_L}{C_\theta}\right\rfloor}\left(\frac{\Lambda}{\mu}\right)^k}, \left\lfloor\frac{\mu V_L}{C_\theta}\right\rfloor < n \leqslant n_\theta^P \end{cases} \tag{D.5}$$

$$\hat{\pi}_\theta^P(n,V_H,\widehat{\sigma}_\theta^i,\widehat{\sigma}_\theta^u) = \begin{cases} \frac{\left(\frac{\Lambda}{\mu}\right)^n}{1+\sum_{k=1}^{k=n_\theta^P}\left(\frac{\Lambda}{\mu}\right)^k + \beta\sum_{k=n_P^*+1}^{k=\left\lfloor\frac{\mu V_H}{C_\theta}\right\rfloor-1}\left(\frac{\Lambda}{\mu}\right)^k}, n \leqslant n_\theta^P \\ \frac{\beta\left(\frac{\Lambda}{\mu}\right)^n}{1+\sum_{k=1}^{k=n_\theta^P}\left(\frac{\Lambda}{\mu}\right)^k + \beta\sum_{k=n_\theta^P+1}^{k=\left\lfloor\frac{\mu V_H}{C_\theta}\right\rfloor}\left(\frac{\Lambda}{\mu}\right)^k}, n_\theta^P < n \leqslant \left\lfloor\frac{\mu V_H}{C_\theta}\right\rfloor \end{cases}. \tag{D.6}$$

For the uninformed customers, if they observe more than $n_\theta^P + 1$ customers awaiting in the system, they would join the queue because $\hat{\pi}_\theta^P(n,V_L,\widehat{\sigma}_\theta^i,\widehat{\sigma}_\theta^u) = 0$, $\forall n > n_\theta^P$. They would infer that with probability one the quality would be high, and this is consistent with the belief under projection. If uninformed customers observe exactly n_θ^P, customers waiting in the system, they would not join the queue simply due to the negative payoffs. If the service quality is high, the informed customers would cross the "hole" for them. If the service quality is low, the informed customers would not cross the "hole" for them. Now, we know that the informed $L-$type customers would not join at $n = n_L^P$, and uninformed customers would never observe more than $n_L^P + 1$ awaiting customers for the low quality service, the queue length would stop at n_L^P. For $H-$type customers however, although the informed $L-$type customers would not join at $n = n_H^P$, the uninformed $L-$type customers would join

nonetheless. Under the projection effects, the $H-$type customers mistakenly believe that the entering customers are of $H-$type and form a biased belief that the service quality must be high with probability one. Therefore, the uninformed $H-$type customers would join until $n = \left\lfloor \frac{\mu V_H}{C_H} \right\rfloor - 1$.

Finally, we need to check that $\forall\theta \in \{H, L\}$, n_θ^P exist and are unique for $\forall\theta \in \{H, L\}$. Let $\frac{\hat{\pi}_\theta^P(n,V_L,\widehat{\sigma}_\theta^i,\widehat{\sigma}_\theta^u)}{\hat{\pi}_\theta^P(n,V_H,\widehat{\sigma}_\theta^i,\widehat{\sigma}_\theta^u)} \equiv l_\theta^P(n,\widehat{\sigma}_\theta^i,\widehat{\sigma}_\theta^u)$ be the estimated likelihood ratio. For any $n_\theta^P \leqslant \left\lfloor \frac{\mu V_H}{C_\theta} \right\rfloor$, the likelihood ratio is well-defined by the limiting distribution of the birth-death process:

$$
\begin{aligned}
l_\theta^P(n_\theta^P,\widehat{\sigma}_\theta^i,\widehat{\sigma}_\theta^u) &= \frac{\left[\frac{(1-\beta)^{n_\theta^P-\left\lfloor\frac{\mu V_L}{C_\theta}\right\rfloor}\left(\frac{\Lambda}{\mu}\right)^n}{1+\sum_{k=1}^{k=\left\lfloor\frac{\mu V_L}{C_\theta}\right\rfloor}\left(\frac{\Lambda}{\mu}\right)^k+\sum_{k=\left\lfloor\frac{\mu V_L}{C_\theta}\right\rfloor+1}^{k=n_\theta^P}(1-\beta)^{k-\left\lfloor\frac{\mu V_L}{C_\theta}\right\rfloor}\left(\frac{\Lambda}{\mu}\right)^k}\right]}{\left[\frac{\left(\frac{\Lambda}{\mu}\right)^{n_\theta^P}}{1+\sum_{k=1}^{k=n_\theta^P}\left(\frac{\Lambda}{\mu}\right)^k+\beta\sum_{k=n_\theta^P+1}^{k=\left\lfloor\frac{\mu V_H}{C_\theta}\right\rfloor}\left(\frac{\Lambda}{\mu}\right)^k}\right]} \\
&= \frac{(1-\beta)^{n_\theta^P-\left\lfloor\frac{\mu V_L}{C_\theta}\right\rfloor}\left[1+\sum_{k=1}^{k=n_\theta^P}\left(\frac{\Lambda}{\mu}\right)^k+\beta\sum_{k=n_\theta^P+1}^{k=\left\lfloor\frac{\mu V_H}{C_\theta}\right\rfloor}\left(\frac{\Lambda}{\mu}\right)^k\right]}{1+\sum_{k=1}^{k=\left\lfloor\frac{\mu V_L}{C_\theta}\right\rfloor}\left(\frac{\Lambda}{\mu}\right)^k+\sum_{k=\left\lfloor\frac{\mu V_L}{C_\theta}\right\rfloor+1}^{k=n_\theta^P}(1-\beta)^{k-\left\lfloor\frac{\mu V_L}{C_\theta}\right\rfloor}\left(\frac{\Lambda}{\mu}\right)^k}.
\end{aligned}
\tag{D.7}
$$

In equilibrium, the anticipated n_θ^P should be consistent, which means:

$$
n_\theta^P = \min\left\{ n \in Z^+ \left| \begin{array}{c} n > \frac{\mu V_H}{C_\theta}\frac{\pi_0}{\pi_0+(1-\pi_0)l_\theta^P(n,\widehat{\sigma}_\theta^i,\widehat{\sigma}_\theta^u)} + \frac{\mu V_L}{C_\theta}\cdot\frac{(1-\pi_0)}{\frac{\pi_0}{l_\theta^P(n,\widehat{\sigma}_\theta^i,\widehat{\sigma}_\theta^u)}+(1-\pi_0)} - 1 \\ \left\lfloor\frac{\mu V_L}{C_H}\right\rfloor \leqslant n \leqslant \left\lfloor\frac{\mu V_H}{C_H}\right\rfloor \end{array} \right.\right\}. \tag{D.8}
$$

For notational simplicity, we define:

$$
\Phi_\theta^P(n) = \frac{1+\sum_{k=1}^{k=n}\left(\frac{\Lambda}{\mu}\right)^k+\beta\sum_{k=n+1}^{k=\left\lfloor\frac{\mu V_H}{C_\theta}\right\rfloor}\left(\frac{\Lambda}{\mu}\right)^k}{1+\sum_{k=1}^{k=\left\lfloor\frac{\mu V_L}{C_\theta}\right\rfloor}\left(\frac{\Lambda}{\mu}\right)^k+\sum_{k=\left\lfloor\frac{\mu V_L}{C_\theta}\right\rfloor+1}^{k=n}(1-\beta)^{k-\left\lfloor\frac{\mu V_L}{C_\theta}\right\rfloor}\left(\frac{\Lambda}{\mu}\right)^k}, \tag{D.9}
$$

$$
N_\theta^P(\Phi_\theta^P) = \min\left\{ n \in Z^+ \left| \begin{array}{c} (1-\beta)^{n-\left\lfloor\frac{\mu V_L}{C_\theta}\right\rfloor}\Phi_\theta^P > \frac{\pi_0}{(1-\pi_0)}\frac{V_H-(n+1)C_\theta/\mu}{(n+1)C_\theta/\mu-V_L} \\ \left\lfloor\frac{\mu V_L}{C_H}\right\rfloor \leqslant n \leqslant \left\lfloor\frac{\mu V_H}{C_H}\right\rfloor \end{array} \right.\right\}. \tag{D.10}
$$

Therefore, the equilibrium n_θ^P are the fixed-points of the following equations system:

$$
\Phi_\theta^P\left(N_\theta^P(\phi_\theta^P)\right) = \phi_\theta^P. \tag{D.11}
$$

To establish the uniqueness of the equilibrium, we are left to study the properties of the two functions $\Phi_\theta^P(n)$ and $N_\theta^P(\Phi_\theta^P)$. First, consider the continuous relaxation of the equations system, i.e., $n \in \mathbb{R}$. $N_\theta^P(\Phi_\theta^P)$ is non-increasing in Φ_θ^P, since whenever $(1-\beta)^{n-\left\lfloor \frac{\mu V_L}{C_\theta} \right\rfloor} \Phi_\theta^P > \frac{\pi_0}{(1-\pi_0)} \frac{V_H-(n+1)C_\theta/\mu}{(n+1)C_\theta/\mu - V_L}$, for $\forall\ \epsilon > 0$, $(1-\beta)^{n-\left\lfloor \frac{\mu V_L}{C_\theta} \right\rfloor}(\Phi_\theta^P + \epsilon) > \frac{\pi_0}{(1-\pi_0)} \frac{V_H-(n+1)C_\theta/\mu}{(n+1)C_\theta/\mu - V_L}$. On the other hand, $\Phi_\theta^P(n)$ is increasing in n due to the Proposition 2 in Debo et al. (2012). Therefore, $\Phi_\theta^P\left(N_\theta^P(\phi_\theta^P)\right)$ is decreasing in ϕ_θ^P. For $\phi_\theta^P = 0$, $\Phi_\theta^P\left(N_\theta^P(\phi_\theta^P)\right)$ is strictly positive. As $\phi_\theta^P \to \infty$, $\Phi_\theta^P\left(N_\theta^P(\phi_\theta^P)\right)$ has unique intersection with the 45° line, provided that $\left\lfloor \frac{\mu V_L}{C_H} \right\rfloor \leqslant N_\theta^P(\phi_\theta^P) \leqslant \left\lfloor \frac{\mu V_H}{C_H} \right\rfloor$. □

Proof of Proposition 27.

The uninformed customers would join if:

$$W_u(n,\theta) = \alpha_\theta(n) W_i(n,\theta,V_H) + (1-\alpha_\theta(n)) W_i(n,\theta,V_L) \geqslant 0, \tag{D.12}$$

which could be rewritten in terms of the cost-benefit analysis:

$$\frac{\pi_0}{\pi_0 + (1-\pi_0) l(n,\widehat{\sigma}_\theta^i,\widehat{\sigma}_\theta^u)} V_H + \frac{(1-\pi_0)}{\pi_0 / l(n,\widehat{\sigma}_\theta^i,\widehat{\sigma}_\theta^u) + (1-\pi_0)} V_L \geqslant \frac{(n+1)C_\theta}{\mu}, \tag{D.13}$$

where $l(n,\widehat{\sigma}_\theta^i,\widehat{\sigma}_\theta^u) \triangleq \frac{\pi(n,V_L,\widehat{\sigma}_\theta^i,\widehat{\sigma}_\theta^u)}{\pi(n,V_H,\widehat{\sigma}_\theta^i,\widehat{\sigma}_\theta^u)}$. Under Bayesian updating, the limiting distribution is determined by the birth-death process. Hence:

$$l(n,\widehat{\sigma}_\theta^i,\widehat{\sigma}_\theta^u) = \begin{cases} \Phi(\vec{n}), n \leqslant \left\lfloor \frac{\mu V_L}{C_H} \right\rfloor \\ \frac{(1-\gamma\beta)^n}{(1-\gamma\beta)^{\left\lfloor \frac{\mu V_L}{C_H} \right\rfloor}} \Phi(\vec{n}), \left\lfloor \frac{\mu V_L}{C_H} \right\rfloor < n \leqslant n_H^1 \\ \frac{(1-\gamma)}{(1-\gamma+\gamma\beta)(1-\gamma\beta)^{\left\lfloor \frac{\mu V_L}{C_H} \right\rfloor + 1}} (1-\gamma\beta)^n \Phi(\vec{n}), n_H^1 < n \leqslant n_H^2 \\ \cdots \\ \frac{(1-\gamma\beta)^{n_H^{|S|} - \left\lfloor \frac{\mu V_L}{C_H} \right\rfloor - |S|} (1-\gamma+\gamma\beta)^{n_H^{|S|} - |S|}}{(1-\gamma)^{n_H^{|S|} - |S|}} \left(\frac{1-\gamma}{1-\gamma+\gamma\beta} \right)^n \Phi(\vec{n}), n_H^{|S|} < n \leqslant \left\lfloor \frac{\mu V_H}{C_H} \right\rfloor \\ \frac{(1-\gamma\beta)^{n_H^{|S|} - \left\lfloor \frac{\mu V_L}{C_H} \right\rfloor - |S|}}{(1-\gamma+\gamma\beta)^{\left\lfloor \frac{\mu V_H}{C_H} \right\rfloor - n_H^{|S|} + |S|} (1-\gamma)^{n_H^{|S|} - |S| - \left\lfloor \frac{\mu V_H}{C_H} \right\rfloor}} \Phi(\vec{n}), \left\lfloor \frac{\mu V_H}{C_H} \right\rfloor < n \leqslant \left\lfloor \frac{\mu V_L}{C_L} \right\rfloor \\ \frac{(1-\gamma\beta)^{n_H^{|S|} - \left\lfloor \frac{\mu V_L}{C_H} \right\rfloor - |S|}}{(1-\beta)^{\left\lfloor \frac{\mu V_L}{C_L} \right\rfloor} (1-\gamma+\gamma\beta)^{\left\lfloor \frac{\mu V_H}{C_H} \right\rfloor - n_H^{|S|} + |S|} (1-\gamma)^{n_H^{|S|} - |S| - \left\lfloor \frac{\mu V_H}{C_H} \right\rfloor}} (1-\beta)^n \Phi(\vec{n}), \left\lfloor \frac{\mu V_L}{C_L} \right\rfloor < n \leqslant n_L^* \\ 0, n_L^* < n \leqslant \left\lfloor \frac{\mu V_H}{C_L} \right\rfloor \end{cases}, \tag{D.14}$$

where,

$$\Phi(\vec{n}) \quad = \quad \Phi(n_H^1, n_H^2, ..., n_H^{|S|}, n_L^*) = \frac{\pi(0,V_L,\widehat{\sigma}_\theta^i,\widehat{\sigma}_\theta^u)}{\pi(0,V_H,\widehat{\sigma}_\theta^i,\widehat{\sigma}_\theta^u)}.$$

For notational simplicity, define "virtual valuation" for joining the service:

$$\hat{V}(n) = \frac{\pi_0}{\pi_0 + (1-\pi_0)l(n, \widehat{\sigma}_\theta^i, \widehat{\sigma}_\theta^u)} V_H + \frac{(1-\pi_0)}{\pi_0 / l(n, \widehat{\sigma}_\theta^i, \widehat{\sigma}_\theta^u) + (1-\pi_0)} V_L. \tag{D.15}$$

By consistency and enforcing the equilibrium such that $(\widehat{\sigma}_\theta^i, \widehat{\sigma}_\theta^u) = (\sigma_\theta^i, \sigma_\theta^u)$, we have:

$$\begin{cases} \hat{V}(n_L^* - 1) \geqslant \frac{n_L^* C_L}{\mu} \\ \hat{V}(n_L^*) < \frac{\left(n_L^*+1\right) C_L}{\mu} \\ \hat{V}(n_L^* + 1) \geqslant \frac{\left(n_L^*+2\right) C_L}{\mu} \end{cases}. \tag{D.16}$$

Since $\hat{V}(n_L^* + 1) = V_H \geqslant \frac{\left(n_L^*+2\right) C_L}{\mu}$, it suffices to require:

$$n_L^* = \min \left\{ n \,\middle|\, \begin{array}{c} \hat{V}(n) < \frac{(n+1)C_\theta}{\mu} \\ \left\lfloor \frac{\mu V_L}{C_L} \right\rfloor < n \leqslant \left\lfloor \frac{\mu V_H}{C_L} \right\rfloor \end{array} \right\}. \tag{D.17}$$

In addition, we need:

$$\begin{cases} \hat{V}(n_H^s - 1) \geqslant \frac{n_H^s C_H}{\mu} \\ \hat{V}(n_H^s) < \frac{\left(n_H^s+1\right) C_H}{\mu} \\ \hat{V}(n_H^s + 1) \geqslant \frac{\left(n_H^s+2\right) C_H}{\mu} \end{cases}, \forall s = 1, 2, ..., |S| - 1, \tag{D.18}$$

and,

$$\begin{cases} \hat{V}(n_H^{|S|} - 1) \geqslant \frac{n_H^{|S|} C_H}{\mu} \\ \hat{V}(n) < \frac{(n+1)C_H}{\mu}, \forall n, n_H^{|S|} \leqslant n \leqslant \left\lfloor \frac{\mu V_H}{C_H} \right\rfloor \end{cases}. \tag{D.19}$$

Define $N(l)$ as the mapping from given likelihood ratio function to $\vec{n}$, satisfying the set of inequalities (D.17)-(D.19). We also know the exact forms of $l(\Phi(\vec{n}))$, and define $L(\vec{n}) = l(\Phi(\vec{n}))$. In equilibrium, we must have $N\left(L\left(\left\{n_H^1, n_H^2, ..., n_H^{|S|}, n_L^*\right\}\right)\right) = \left\{n_H^1, n_H^2, ..., n_H^{|S|}, n_L^*\right\}$. The proof on the uniqueness of n_L^* is a repetition to the proof of Proposition 28. □

Proof of Proposition 29. For the informed customers, their strategy profiles are exactly the same as under the projection effects.

The uninformed customers would join if

$$W_u(n, \theta, \hat{\gamma}_\theta^R) = \hat{\alpha}_\theta^R(n) W_i(n, \theta, V_H) + (1 - \hat{\alpha}_\theta^R(n)) W_i(n, \theta, V_L) \geqslant 0, \tag{D.20}$$

i.e.,

$$n \leqslant \frac{\mu \left[V_H - \left(1 - \hat{\alpha}_\theta^R(n)\right)(V_H - V_L)\right]}{C_\theta} - 1. \tag{D.21}$$

The uninformed $L-$type customers would believe that no one would join for $n \geqslant \left\lfloor \frac{\mu V_H}{C_H} \right\rfloor$ because they mistakenly think that everybody else than themselves would be of $H-$type.

However, the informed $L-$type customers would join for $n \leqslant \left\lfloor \frac{\mu V_L}{C_L} \right\rfloor - 1$ regardless of service quality. By the off-equilibrium belief assumptions, the uninformed $L-$type customers would believe that the service is of high quality with probability one and they join for $n \leqslant \left\lfloor \frac{\mu V_H}{C_L} \right\rfloor - 1$.

For the uninformed $H-$type customers, since they would stop joining at $\frac{\mu\left[V_H - \left(1-\hat{\alpha}_\theta^R(n)\right)(V_H - V_L)\right]}{C_H} - 1$, and:

$$\frac{\mu\left[V_H - \left(1-\hat{\alpha}_\theta^R(n)\right)(V_H - V_L)\right]}{C_H} - 1 < \frac{\mu V_H}{C_H} - 1 \leqslant \left\lfloor \frac{\mu V_H}{C_H} \right\rfloor - 1, \tag{D.22}$$

$$\begin{aligned} \frac{\mu\left[V_H - \left(1-\hat{\alpha}_\theta^R(n)\right)(V_H - V_L)\right]}{C_H} - 1 &= \frac{\mu\left[V_L + \hat{\alpha}_\theta^R(n)(V_H - V_L)\right]}{C_H} - 1 \\ &> \frac{\mu V_L}{C_H} - 1 \geqslant \left\lfloor \frac{\mu V_L}{C_H} \right\rfloor - 1. \end{aligned} \tag{D.23}$$

We know that there would exist a threshold $n_H^R \in \left[\left\lfloor \frac{\mu V_L}{C_H} \right\rfloor - 1, \left\lfloor \frac{\mu V_H}{C_H} \right\rfloor - 1\right]$ such that $\sigma_H^u(V_\varphi, R, n) = 0$ for $n \geqslant n_H^R$. We need to that $\sigma_H^u(V_\varphi, R, n) = 1$ for $n < n_H^R$, or there is no other "hole". To see this, we have:

$$\begin{aligned} \frac{\mu\left[V_H - \left(1-\hat{\alpha}_\theta^R(n)\right)(V_H - V_L)\right]}{C_H} &= \frac{\mu V_H}{C_H} \cdot \frac{\pi_0}{\pi_0 + (1-\pi_0)\frac{\hat{\pi}_H^R(n, V_L, \widehat{\sigma}_\theta^i, \widehat{\sigma}_\theta^u)}{\hat{\pi}_H^R(n, V_H, \widehat{\sigma}_\theta^i, \widehat{\sigma}_\theta^u)}} \\ &\quad + \frac{\mu V_L}{C_H} \cdot \frac{(1-\pi_0)}{\pi_0 \frac{\hat{\pi}_H^R(n, V_H, \widehat{\sigma}_\theta^i, \widehat{\sigma}_\theta^u)}{\hat{\pi}_H^R(n, V_L, \widehat{\sigma}_\theta^i, \widehat{\sigma}_\theta^u)} + (1-\pi_0)}, \end{aligned}$$

where the putatively correct Bayesian updating is determined by the birth-death process.

Let $\frac{\hat{\pi}_H^R(n, V_L, \widehat{\sigma}_\theta^i, \widehat{\sigma}_\theta^u)}{\hat{\pi}_H^R(n, V_H, \widehat{\sigma}_\theta^i, \widehat{\sigma}_\theta^u)} \equiv l_H^R(\widehat{\sigma}_\theta^i, \widehat{\sigma}_\theta^u)$ be the estimated likelihood ratio, which is well defined on the whole support $n \in \left[0, \left\lfloor \frac{\mu V_H}{C_H} \right\rfloor - 1\right]$ that is relevant for the H-type customers' decision making:

$$l_H^R(n, \widehat{\sigma}_\theta^i, \widehat{\sigma}_\theta^u) = \frac{\left[\frac{\left(\frac{\Lambda}{\mu}\right)^n}{1+\sum_{k=1}^{k=\left\lfloor \frac{\mu V_L}{C_L} \right\rfloor - 1}\left(\frac{\Lambda}{\mu}\right)^k + \sum_{k=\left\lfloor \frac{\mu V_L}{C_L} \right\rfloor}^{k=\left\lfloor \frac{\mu V_H}{C_L} \right\rfloor - 1}(1-\beta)^{k-\left\lfloor \frac{\mu V_L}{C_L} \right\rfloor + 1}\left(\frac{\Lambda}{\mu}\right)^k}\right]}{\left[\frac{\left(\frac{\Lambda}{\mu}\right)^n}{1+\sum_{k=1}^{k=\left\lfloor \frac{\mu V_H}{C_L} \right\rfloor - 1}\left(\frac{\Lambda}{\mu}\right)^k}\right]} \tag{D.24}$$

$$= \frac{1 + \sum_{k=1}^{k=\left\lfloor \frac{\mu V_H}{C_L} \right\rfloor - 1}\left(\frac{\Lambda}{\mu}\right)^k}{1 + \sum_{k=1}^{k=\left\lfloor \frac{\mu V_L}{C_L} \right\rfloor - 1}\left(\frac{\Lambda}{\mu}\right)^k + \sum_{k=\left\lfloor \frac{\mu V_L}{C_L} \right\rfloor}^{k=\left\lfloor \frac{\mu V_H}{C_L} \right\rfloor - 1}(1-\beta)^{k-\left\lfloor \frac{\mu V_L}{C_L} \right\rfloor + 1}\left(\frac{\Lambda}{\mu}\right)^k}, \tag{D.25}$$

which is independent of the queue length n. Hence,

$$n_H^R = \left\lfloor \frac{\mu V_H}{C_H} \cdot \frac{\pi_0}{\pi_0 + (1-\pi_0) l_H^R} + \frac{\mu V_L}{C_H} \cdot \frac{(1-\pi_0) l_H^R}{\pi_0 + (1-\pi_0) l_H^R} \right\rfloor - 1, \tag{D.26}$$

which is a constant threshold and $n_H^R \in \left[\left\lfloor \frac{\mu V_L}{C_H} \right\rfloor - 1, \left\lfloor \frac{\mu V_H}{C_H} \right\rfloor - 1 \right]$. $\square$

Proof of Proposition 32: We can plug in the equilibrium joining strategies into the limiting distributions given by equations (D.5), (D.6), and etc. The number of customers in the system under projection effects follows birth-death process, with the limiting distribution contingent on the actual service quality. Thus, with such closed-form expressions, we can compare the arrival rates or the likelihood ratios directly. By some algebra which we omitted here, it is straightforward to see that $\mathcal{Q}(HP, LR, V_\varphi) \succeq_{lr} \mathcal{Q}(HP, LP, V_\varphi)$, $\mathcal{Q}(HP, LR, V_\varphi) \succeq_{lr} \mathcal{Q}(HR, LR, V_\varphi) \succeq_{lr} \mathcal{Q}(HR, LP, V_\varphi)$, provided that $n_H^R \leq n_H^P - 1$. In what follows, we show that this sufficient (although not necessary) condition is satisfied.

Define the constant C_0 as:

$$C_0 = \frac{\left[1 + \sum_{k=1}^{k=\left\lfloor \frac{\mu V_H}{C_L} \right\rfloor - 1} \left(\frac{\Lambda}{\mu}\right)^k\right] \left[1 + \sum_{k=1}^{k=\left\lfloor \frac{\mu V_L}{C_H} \right\rfloor} \left(\frac{\Lambda}{\mu}\right)^k + \sum_{k=\left\lfloor \frac{\mu V_L}{C_H} \right\rfloor + 1}^{k=\left\lfloor \frac{\mu V_H}{C_H} \right\rfloor} (1-\beta)^{k-\left\lfloor \frac{\mu V_L}{C_H} \right\rfloor} \left(\frac{\Lambda}{\mu}\right)^k\right]}{\left[1 + \sum_{k=1}^{k=\left\lfloor \frac{\mu V_L}{C_L} \right\rfloor - 1} \left(\frac{\Lambda}{\mu}\right)^k + \sum_{k=\left\lfloor \frac{\mu V_L}{C_L} \right\rfloor}^{k=\left\lfloor \frac{\mu V_H}{C_L} \right\rfloor - 1} (1-\beta)^{k-\left\lfloor \frac{\mu V_L}{C_L} \right\rfloor + 1} \left(\frac{\Lambda}{\mu}\right)^k\right] \left[1 + \sum_{k=1}^{k=\left\lfloor \frac{\mu V_H}{C_H} \right\rfloor} \left(\frac{\Lambda}{\mu}\right)^k\right]}. \tag{D.27}$$

The fact that $n_H^P \geq \left\lfloor \frac{\mu V_L}{C_H} \right\rfloor + 1 + \frac{\log(C_0)}{\log(1-\beta)}$ indicates:

$$\begin{aligned} l_H^R &= \frac{1 + \sum_{k=1}^{k=\left\lfloor \frac{\mu V_H}{C_L} \right\rfloor - 1} \left(\frac{\Lambda}{\mu}\right)^k}{1 + \sum_{k=1}^{k=\left\lfloor \frac{\mu V_L}{C_L} \right\rfloor - 1} \left(\frac{\Lambda}{\mu}\right)^k + \sum_{k=\left\lfloor \frac{\mu V_L}{C_L} \right\rfloor}^{k=\left\lfloor \frac{\mu V_H}{C_L} \right\rfloor - 1} (1-\beta)^{k-\left\lfloor \frac{\mu V_L}{C_L} \right\rfloor + 1} \left(\frac{\Lambda}{\mu}\right)^k} \\ &\geq \frac{(1-\beta)^{n_H^P - 1 - \left\lfloor \frac{\mu V_L}{C_H} \right\rfloor} \left[1 + \sum_{k=1}^{k=\left\lfloor \frac{\mu V_H}{C_H} \right\rfloor} \left(\frac{\Lambda}{\mu}\right)^k\right]}{1 + \sum_{k=1}^{k=\left\lfloor \frac{\mu V_L}{C_H} \right\rfloor} \left(\frac{\Lambda}{\mu}\right)^k + \sum_{k=\left\lfloor \frac{\mu V_L}{C_H} \right\rfloor + 1}^{k=\left\lfloor \frac{\mu V_H}{C_H} \right\rfloor} (1-\beta)^{k-\left\lfloor \frac{\mu V_L}{C_H} \right\rfloor} \left(\frac{\Lambda}{\mu}\right)^k} = \Phi_H^P\left(\left\lfloor \frac{\mu V_H}{C_H} \right\rfloor + 1\right). \end{aligned}$$

Since $\Phi_H^P(n)$ is increasing in n, we have $\Phi_H^P(n_H^P - 1) < \Phi_H^P(\left\lfloor \frac{\mu V_H}{C_H} \right\rfloor + 1)$, and thus $l_H^R \geq (1-\beta)^{n_H^P - 1 - \left\lfloor \frac{\mu V_L}{C_H} \right\rfloor} \Phi_H^P(n_H^P - 1)$. This suggest that $n_H^R \leq n_H^P - 1$.

To compare the queue lengths in terms of stochastic dominance, we need the following well-known lemma:

Lemma 7 $\mathcal{X} \succeq_{lr} \mathcal{Y} \Rightarrow \mathcal{X} \succeq_{st} \mathcal{Y}$.

In terms of the expectation measure, we need the following lemma:

Lemma 8 *$\mathcal{X} \succeq_{st} \mathcal{Y}$, if and only if $\mathrm{E}h(\mathcal{X}) \geq \mathrm{E}h(\mathcal{Y})$, for any nondecreasing function $h(\cdot)$.*

Therefore, the same ordering holds in terms of the first-order stochastic dominance as well as the expected queue length. □

Bibliography

Acemoglu, D., K. Bimpikis, and A. Ozdaglar (2014). Dynamics of information exchange in endogenous social networks. *Theoretical Economics 9*(1), 41–97.

Adeyi, O. and R. Atun (2010). Universal access to malaria medicines: Innovation in financing and delivery. *The lancet 376*(9755), 1869–1871.

Afeche, P. (2004). Incentive-compatible revenue management in queueing systems: Optimal strategic idleness and other delaying tactics. *Unpublished Manuscript*.

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

Afeche, P. and J. M. Pavlin (2015). Optimal price/lead-time menus for queues with customer choice: Segmentation, pooling, and strategic delay. *Management Science (forthcoming)*.

Ahlin, C. and P. D. Ahlin (2013). Product differentiation under congestion: Hotelling was right. *Economic Inquiry 51*(3), 1750–1763.

Akgun, O. T., R. Righter, and R. Wolff (2012). Understanding the marginal impact of customer flexibility. *Queueing Systems 71*(1-2), 5–23.

Akgun, O. T., R. Righter, R. Wolff, et al. (2011). Multiple-server system with flexible arrivals. *Advances in Applied Probability 43*(4), 985–1004.

Allon, G. and A. Federgruen (2007). Competition in service industries. *Operations Research 55*(1), 37–55.

Alptekinoglu, A. and C. J. Corbett (2010). Leadtime-variety tradeoff in product differentiation. *Manufacturing & Service Operations Management 12*(4), 569–582.

An, J., S.-H. Cho, and C. S. Tang (2015). Aggregating smallholder farmers in emerging economies. *Production and Operations Management*, forthcoming.

Ashraf, N., J. Berry, and J. M. Shapiro (2010). Can higher prices stimulate product use? Evidence from a field experiment in Zambia. *American Economic Review 100*, 2383–2413.

Ashraf, N., D. Karlan, and W. Yin (2006). Tying odysseus to the mast: Evidence from a commitment savings product in the Philippines. *The Quarterly Journal of Economics*, 635–672.

Atkin, D., A. Chaudhry, S. Chaudry, A. Khandelwal, and E. A. Verhoogen (2015). Organizational barriers to technology adoption: Evidence from soccer-ball producers in pakistan. *NBER Working Paper* (w21417).

Bala, V. and S. Goyal (2000). A noncooperative model of network formation. *Econometrica 68*(5), 1181–1229.

Bandiera, O. and I. Rasul (2006). Social networks and technology adoption in northern Mozambique. *The Economic Journal 116*(514), 869–902.

Banerjee, A. and S. Mullainathan (2010). The shape of temptation: Implications for the economic lives of the poor. Technical report, National Bureau of Economic Research.

Bassamboo, A., R. S. Randhawa, and J. A. V. Mieghem (2012). A little flexibility is all you need: on the asymptotic value of flexible capacity in parallel queuing systems. *Operations Research 60*(6), 1423–1435.

Bernheim, B. D., D. Ray, and S. Yeltekin (2013). Poverty and self-control. Technical report, National Bureau of Economic Research.

Bloch, F. and B. Dutta (2009). Communication networks with endogenous link strength. *Games and Economic Behavior 66*(1), 39–56.

Brander, J. A. and B. . J. Spencer (1985). Export subsidies and international market share rivalry. *Journal of International Economics 18*, 83–100.

Bryan, G., D. Karlan, and S. Nelson (2010). Commitment devices. *Annual Review of Economics 2*(1), 671–698.

Cachon, G. P. and P. T. Harker (2002). Competition and outsourcing with scale economies. *Management Science 48*(10), 1314–1333.

Carter, C. A. and D. MacLaren (1994). Alternative oligopolistic structures in international commodity markets: price or quantity competition? *Working Paper, Department of Agriculture, The University of Melbourne*.

Cattani, K. D., E. Dahan, and G. M. Schmidt (2010). Lowest cost may not lower total cost: Using spackling to smooth mass-customized production. *Production and Operations Management 19*(5), 531–545.

Çelen, B. and S. Kariv (2004). Observational learning under imperfect information. *Games and Economic Behavior 47*(1), 72–86.

Chamley, C. (2004). *Rational herds: Economic models of social learning*. Cambridge University Press.

Chen, Y.-J., J. G. Shanthikumar, and Z.-J. M. Shen (2013). Training, production, and channel separation in ITC's e-Choupal network. *Production and Operations Management 22*(2), 348–364.

Chen, Y.-J., J. G. Shanthikumar, and Z.-J. M. Shen (2014). Incentive for peer-to-peer knowledge sharing among farmers in developing economies. *Production and Operations Management*, forthcoming.

Chen, Y.-J. and C. S. Tang (2015). The economic value of market information for farmers in developing economies. *Production and Operations Management*, forthcoming.

Cohen, J. and P. Dupas (2010). Free distribution or cost-sharing? Evidence from a randomized malaria prevention experiment. *Quarterly Journal of Economics 125*(1).

Cole, S., X. Giné, J. Tobacman, R. Townsend, P. Topalova, and J. Vickery (2013). Barriers to household risk management: Evidence from India. *American Economic Journal. Applied economics 5*(1), 104.

Colombo, L., G. Femminis, and A. Pavan (2014). Information acquisition and welfare. *The Review of Economic Studies*, forthcoming.

Conley, T. G. and C. R. Udry (2010). Learning about a new technology: Pineapple in Ghana. *American Economic Review 100*(1), 35–69.

Cui, S. and S. K. Veeraraghavan (2014). Blind queues: The impact of consumer beliefs on revenues and congestion. *Available at SSRN 2196817*.

Cui, T., Y.-J. Chen, and Z.-J. M. Shen (2009). Optimal pricing, scheduling, and admission control for queueing systems under information asymmetry. *Working paper, University of California at Berkeley*.

Currarini, S. and F. Feri (2014). Information sharing networks in linear quadratic games. *International Journal of Game Theory*, 1–32.

Dasgupta, A. et al. (2000). Social learning with payoff complementarities. *London School of Economics 25*.

Debo, L. and S. Veeraraghavan (2014). Equilibrium in queues under unknown service times and service value. *Operations Research 62*(1), 38–57.

Debo, L. G., C. Parlour, and U. Rajan (2012). Signaling quality via queues. *Management Science 58*(5), 876–891.

DeMarzo, P. M., D. Vayanos, and J. Zwiebel (2003). Persuasion bias, social influence, and uni-dimensional opinions. *Quarterly Journal of Economics 118*(3), 909–968.

Deodhar, S. Y. and I. M. Sheldon (1996). Estimation of imperfect competition in food marketing: A dynamic analysis of the German banana market. *Journal of Food Distribution Research 27*, 1–10.

Devaux, A., D. Horton, C. Velasco, G. Thiele, G. Lpez, T. Bernet, I. Reinoso, and M. Ordinola (2009). Collective action for market chain innovation in the Andes. *Food Policy 34*(1), 31 – 38.

Duflo, E., M. Kremer, and J. Robinson (2008). How high are rates of return to fertilizer? Evidence from field experiments in Kenya. *The American Economic Review*, 482–488.

Duflo, E., M. Kremer, and J. Robinson (2011). Nudging farmers to use fertilizer: Theory and experimental evidence from Kenya. *American Economic Review 101*, 2350–2390.

Dugger, C. W. (2007). Ending famine, simply by ignoring the experts. *New York Times 2*(12), 07.

Dupas, P. (2014). Short-run subsidies and long-run adoption of new health products: Evidence from a field experiment. *Econometrica: Journal of the Econometric Society 82*(1), 197.

Dupas, P. and J. Robinson (2011). Why don't the poor save more? Evidence from health savings experiments. Technical report, National Bureau of Economic Research.

Eliaz, K. and R. Spiegler (2008). Consumer optimism and price discrimination. *Theoretical Economics 3*(4), 459–497.

Eyster, E., A. Galeotti, N. Kartik, and M. Rabin (2014). Congested observational learning. *Games and Economic Behavior 87*, 519–538.

Eyster, E. and M. Rabin (2005). Cursed equilibrium. *Econometrica 73*(5), 1623–1672.

Eyster, E. and M. Rabin (2010). Naive herding in rich-information settings. *American Economic Journal: Microeconomics 2*(4), 221–243.

Fang, H. and Y. Wang (2015). Estimating dynamic discrete choice models with hyperbolic discounting, with an application to mammography decisions. *International Economic Review 56*(2), 565–596.

Fischer, E. and M. Qaim (2012). Linking smallholders to markets: Determinants and impacts of farmer collective action in Kenya. *World Development 40*(6), 1255 – 1268.

Frable, D. E. (1993). Being and feeling unique: Statistical deviance and psychological marginality. *Journal of Personality 61*(1), 85–110.

Gagnon-Bartsch, T. (2014). Taste projection in a model of social learning. *Working paper*.

Gal-Or, E. (1985). Information sharing in oligopoly. *Econometrica 53*(2), pp. 329–343.

Galeotti, A. and S. Goyal (2010). The law of the few. *The American Economic Review*, 1468–1492.

Gans, N., G. Koole, and A. Mandelbaum (2003). Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management 5*(2), 79–141.

Garnevska, E., G. Liu, and N. M. Shadbolt (2011). Factors for successful development of farmer cooperatives in northwest China. *International Food and Agribusiness Management Review 14*(4).

Giné, X. and D. Yang (2009). Insurance, credit, and technology adoption: Field experimental evidencefrom Malawi. *Journal of Development Economics 89*(1), 1–11.

Golovina, S., J. Nilsson, et al. (2009). Russian agricultural producers' views of top-down organized cooperatives. *Journal of rural cooperation 37*(2), 225.

Goyal, S. and S. Joshi (2003). Networks of collaboration in oligopoly. *Games and Economic Behavior 43*(1), 57 – 85.

Greenwood, B. M., K. Bojang, C. J. Whitty, and G. A. Targett (2005). Malaria. *The Lancet 365*(9469), 1487 – 1498.

Guiteras, R., D. I. Levine, T. Polley, and B. Quistorff (2013). Credit constraints, present bias and investment in health: Evidence from micropayments for clean water in Dhaka. *Unpublished*.

Ha, A. Y. (1998). Incentive-compatible pricing for a service facility with joint production and congestion externalities. *Management Science 44*(12-part-1), 1623–1636.

Ha, A. Y. (2001). Optimal pricing that coordinates queues with customer-chosen service requirements. *Management Science 47*(7), 915–930.

Hassan, S. E.-D. H., E. M. Malik, S. I. Okoued, and E. M. Eltayeb (2008). Retention and efficacy of long-lasting insecticide-treated nets distributed in eastern Sudan: A two-step community-based study. *Malarial Journal 7*(1), 85.

Hassin, R. and M. Haviv (2003). *To queue or not to queue: Equilibrium behavior in queueing systems*, Volume 59. Springer Science & Business Media.

He, Y.-T. and D. G. Down (2009). On accommodating customer flexibility in service systems. *INFOR: Information Systems and Operational Research 47*(4), 289–295.

Hellin, J., M. Lundy, and M. Meijer (2009). Farmer organization, collective action and market access in Meso-America. *Food Policy 34*(1), 16 – 22.

Hellwig, C. and L. Veldkamp (2009). Knowing what others know: Coordination motives in information acquisition. *The Review of Economic Studies 76*(1), 223–251.

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

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

Jackson, M. O. (2010). *Social and economic networks*. Princeton University Press.

Jia, X. and J. Huang (2011). Contractual arrangements between farmer cooperatives and buyers in China. *Food Policy 36*(5), 656 – 666.

Jouini, O., Y. Dallery, and R. Nait-Abdallah (2008). Analysis of the impact of team-based organizations in call center management. *Management Science 54*(2), 400–414.

Kaganzi, E., S. Ferris, J. Barham, A. Abenakyo, P. Sanginga, and J. Njuki (2009). Sustaining linkages to high value markets through collective action in Uganda. *Food Policy 34*(1), 23 – 30.

Kessing, S. G. and R. Nuscheler (2006). Monopoly pricing with negative network effects: The case of vaccines. *European Economic Review 50*(4), 1061–1069.

Koenig, M. (2012). The formation of networks with local spillovers and limited observability. Technical report, Stanford Institute for Economic Policy Research.

Kohlberg, E. (1983). Equilibrium store locations when consumers minimize travel time plus waiting time. *Economics Letters 11*(3), 211–216.

Kremer, M. and C. M. Snyder (2003). Why are drugs more profitable than vaccines? Technical report, National Bureau of Economic Research.

Lee, H. L. and C.-Y. Lee (2007). *Building supply chain excellence in emerging economies*, Volume 98. Springer Science & Business Media.

Lee, Y.-J. (2014). Information sharing networks in oligopoly. *The Korean Economic Review 30*(1), 41–66.

Levine, D. I., T. Beltramo, G. Blalock, and C. Cotterman (2012). What impedes efficient adoption of products? Evidence from randomized variation in sales offers for improved cookstoves in Uganda.

Li, L. (2002). Information sharing in a supply chain with horizontal competition. *Management Science 48*(9), 1196–1212.

Li, L. and H. Zhang (2008). Confidentiality and information sharing in supply chain coordination. *Management Science 54*(8), 1467–1481.

Madarász, K. (2011). Information projection: Model and applications. *The Review of Economic Studies*, 961–985.

Maglaras, C., J. Yao, and A. Zeevi (2013). Optimal price and delay differentiation in queueing systems. *Available at SSRN 2297042*.

Marks, G. and N. Miller (1987). Ten years of research on the false-consensus effect: An empirical and theoretical review. *Psychological Bulletin 102*(1), 72.

Mendelson, H. and A. K. Parlaktürk (2008). Product-line competition: Customization vs. proliferation. *Management Science 54*(12), 2039–2053.

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

Michelson, H., T. Reardon, and F. Perez (2012). Small farmers and big retail: Trade-offs of supplying supermarkets in Nicaragua. *World Development 40*(2), 342 – 354.

Miguel, E. and M. Kremer (2004). Worms: identifying impacts on education and health in the presence of treatment externalities. *Econometrica*, 159–217.

Mobarak, A. M., P. Dwivedi, R. Bailis, L. Hildemann, and G. Miller (2012). Low demand for nontraditional cookstove technologies. *Proceedings of the National Academy of Sciences 109*(27), 10815–10820.

Monasch, R., A. Reinisch, R. W. Steketee, E. L. Korenromp, D. Alnwick, and Y. Bergevin (2004). Child coverage with mosquito nets and malaria treatment from population-based surveys in African countries: A baseline for monitoring progress in roll back malaria. *The American Journal of Tropical Medicine and Hygiene 71*(2 suppl), 232–238.

Moorthy, S. and K. Srinivasan (1995). Signaling quality with a money-back guarantee: The role of transaction costs. *Marketing Science 14*(4), 442–466.

Moorthy, V. S., M. F. Good, and A. V. Hill (2004). Malaria vaccine developments. *The Lancet 363*(9403), 150–156.

Morris, A., A. Ward, B. Moonen, O. Sabot, and J. M. Cohen (2015). Price subsidies increase the use of private sector acts: Evidence from a systematic review. *Health Policy and Planning 30*(3), 397–405.

Morris, S. and H. S. Shin (2002). Social value of public information. *The American Economic Review 92*(5), 1521–1534.

Moustier, P., P. T. G. Tam, D. T. Anh, V. T. Binh, and N. T. T. Loc (2010). The role of farmer organizations in supplying supermarkets with quality food in Vietnam. *Food Policy 35*(1), 69 – 78.

Myatt, D. P. and C. Wallace (2012). Endogenous information acquisition in coordination games. *The Review of Economic Studies 79*(1), 340–374.

Myerson, R. B. (1991). Game theory: analysis of conflict. *Harvard University*.

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

Nieuwoudt, W. L. (1987). Allocation of beef permits and quotas. *South African Journal of Economics 55*(3), 182–187.

O'Donoghue, T. and M. Rabin (2001). Choice and procrastination. *Quarterly Journal of Economics*, 121–160.

O'Donoghue, T. and M. Rabin (2006). Optimal sin taxes. *Journal of Public Economics 90*(1011), 1825 – 1849.

Osborne, T. (2005). Imperfect competition in agricultural markets: evidence from Ethiopia. *Journal of Development Economics 76*(2), 405–428.

Parlaktürk, A. K. (2012). The value of product variety when selling to strategic consumers. *Manufacturing & Service Operations Management 14*(3), 371–385.

Qiao, H. Z. L. and S. Yu (2013). Collective actions of small farm households in big markets: Ruoheng farmer watermelon cooperative in China. *Linking Smallholder Producers to Modern Agri-Food Chains: Case Studies from South Asia, Southeast Asia and China 1*.

Rabin, M. and J. L. Schrag (1999). First impressions matter: A model of confirmatory bias. *Quarterly journal of Economics*, 37–82.

Raith, M. (1996). A general model of information sharing in oligopoly. *Journal of Economic Theory 71*(1), 260 – 288.

Rodriguez, W. and K. ole-MoiYoi (2011). *Building local capacity for health commodity manufacturing: A to Z Textile Mills LTD.* GHD-009. Boston, MA: Harvard Business Publishing.

Ross, L., D. Greene, and P. House (1977). The false consensus effect: An egocentric bias in social perception and attribution processes. *Journal of experimental social psychology 13*(3), 279–301.

Shanthikumar, J. G. and D. D. Yao (1992). Multiclass queueing systems: Polymatroidal structure and optimal scheduling control. *Operations Research 40*(3-supplement-2), S293–S299.

Sharma, V. P. and H. Thaker (2010). Fertiliser subsidy in India: Who are the beneficiaries? *Economic & Political Weekly 45*(12), 69.

Shin, H. and T. I. Tunca (2010). Do firms invest in forecasting efficiently? The effect of competition on demand forecast investments and supply chain coordination. *Operations Research 58*(6), 1592–1610.

Smith, D., J. Dushoff, R. Snow, and S. Hay (2005). The entomological inoculation rate and plasmodium falciparum infection in African children. *Nature 438*(7067), 492–495.

So, K. C. (2000). Price and time competition for service delivery. *Manufacturing & Service Operations Management 2*(4), 392–409.

Sodhi, M. S. and C. S. Tang (2014). Supply-chain research opportunities with the poor as suppliers or distributors in developing countries. *Production and Operations Management 23*(9), 1483–1494.

Suri, T. (2011). Selection and comparative advantage in technology adoption. *Econometrica 79*(1), 159–209.

Tang, C. S., Y. Wang, and M. Zhao (2014). The implications of utilizing market information and adopting agricultural advice for farmers in developing economies. *Production and Operations Management*, forthcoming.

Tarozzi, A. and A. Mahajan (2011). Time inconsistency, expectations and technology adoption: The case of insecticide treated nets. *Economic Research Initiatives at Duke (ERID) Working Paper* (105).

Tarozzi, A., A. Mahajan, B. Blackburn, D. Kopf, L. Krishnan, and J. Yoong (2014). Micro-loans, insecticide-treated bednets, and malaria: Evidence from a randomized controlled trial in Orissa, India. *American Economic Review 104*(7), 1909–41.

Taylor, T. A. and W. Xiao (2014). Subsidizing the distribution channel: Donor funding to improve the availability of malaria drugs. *Management Science 60*(10), 2461–2477.

Tekin, E., W. J. Hopp, and M. P. Van Oyen (2009). Pooling strategies for call center agent cross-training. *IIE Transactions 41*(6), 546–561.

Trebbin, A. and M. Hassler (2012). Farmers producer companies in India: A new concept for collective action? *Environment and Planning-Part A 44*(2), 411.

Veeraraghavan, S. and L. Debo (2009). Joining longer queues: Information externalities in queue choice. *Manufacturing & Service Operations Management 11*(4), 543–562.

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

Vives, X. (1984). Duopoly information equilibrium: Cournot and Bertrand. *Journal of Economic Theory 34*(1), 71–94.

Vives, X. (1988). Aggregation of information in large Cournot markets. *Econometrica*, 851–876.

Whitty, C. J., R. Allan, V. Wiseman, S. Ochola, M. V. Nakyanzi-Mugisha, B. Vonhm, M. Mwita, C. Miaka, A. Oloo, Z. Premji, et al. (2004). Averting a malaria disaster in Africa: Where does the buck stop? *Bulletin of the World Health Organization 82*(5), 381–384.

Wolff, R. W. (1982). Poisson arrivals see time averages. *Operations Research 30*(2), 223–231.

Xia, N. and S. Rajagopalan (2009). Standard vs. custom products: variety, lead time, and price competition. *Marketing Science 28*(5), 887–900.

Xie, J. and S. M. Shugan (2001). Electronic tickets, smart cards, and online prepayments: When and how to advance sell. *Marketing Science 20*(3), 219–243.

Yahalom, T., J. M. Harrison, and S. Kumar (2006). Designing and pricing incentive compatible grades of service in queueing systems. *Working Paper, Graduate School of Business, Stanford University*.

Yang, L., F. De Véricourt, and P. Sun (2013). Time-based competition with benchmark effects. *Manufacturing & Service Operations Management 16*(1), 119–132.

Yang, L., P. Guo, and Y. Wang (2014). Service pricing with loss averse customers. *Available at SSRN 2418303*.

Zhao, X., K. E. Stecke, and A. Prasad (2012). Lead time and price quotation mode selection: Uniform or differentiated? *Production and Operations Management 21*(1), 177–193.

Zhou, J. and Y.-J. Chen (2015). Key leaders in social networks. *Journal of Economic Theory 157*, 212–235.