Ordering Flexibility, Delay Information Heterogeneity, and Time-of-Service Preferences in Operations Management

by

Yang Li

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy
Graduate Department of Rotman School of Management
University of Toronto

Abstract

Ordering Flexibility, Delay Information Heterogeneity, and Time-of-Service Preferences in Operations Management

Yang Li
Doctor of Philosophy
Graduate Department of Rotman School of Management
University of Toronto
2016

This thesis studies three problems in operations management related to the reorder flexibility in supply chain management and operation strategies in service industry.

In Chapter 2, I investigate the value of reorder flexibility under price competition. Although the supply chain literature shows that reorder flexibility increases profits even under quality competition, I show that price competition, arguably a more appropriate price formation model in the presence of reorder flexibility, may yield *opposite* results: (i) Reorder flexibility may *increase* competing firms' initial orders. (ii) Reorder flexibility *hurts profits except* if it reduces initial orders and in addition, demand variability is moderate, reordering is sufficiently inexpensive, and products are sufficiently differentiated. (iii) Firms can avoid the downside of reorder flexibility only in some cases where it hurts profits. In others, firms are trapped in a prisoner's dilemma, whereby *reorder flexibility is the dominant strategy even though it hurts their profits.*

In Chapter 3, I study how the growing prevalence of real-time delay information affects service system performances, i.e., throughput and social welfare. I consider an M/M/1 system with two streams of customers, one informed about real-time delay and one uninformed. I show that the impacts of growing information prevalence on system performance measures are determined by the equilibrium joining behavior of *uninformed* customers. Moreover, throughput and social welfare can be *unimodal* in the fraction of informed customers. In other words, some amount of information heterogeneity in the population can lead to strictly more efficient outcomes than information homogeneity,

under which the queue length is either visible or invisible to all customers.

In the last chapter, I consider the capacity allocation problem when customer arrivals are endogenously determined by system intertemporal delays. I show that the socially optimal solution is in general not aligned with customer self-interested incentives, since customers do not take into account the provider's capacity cost when choosing their time-of-service (TOS). Therefore, in order to retain the incentive compatibility of the socially optimal solution, the provider has to charge for TOS such that price differences across time periods equal the corresponding capacity cost differences.

Acknowledgements

First of all, I am deeply indebted to my thesis supervisors, Professor Philipp Afèche and Professor Ming Hu, for their support and patience throughout my Ph.D. study. This thesis would not have been possible without their continued guidance. Philipp's invaluable vision and suggestions lead me through many dead-ends of my research. I have also learned from him how to be a rigorous researcher with critical thinking. Ming's enthusiasm towards research inspires me to explore a wide range of topics in operations management. I appreciate that he spends countless hours discussing with me. His intelligence also helps me overcome many technical difficulties.

I would like to express my sincere gratitude to Professor Dmitry Krass, who offered me the opportunity to pursue my Ph.D. and serves on my supervisory committee. I am also grateful to Professor Laurens Debo for kindly accepting to be my thesis external appraiser.

I would also like to thank other faculty members in the Operations Management group: Oded Berman, Opher Baron, Joseph Milner, Azarakhshand Malekian, and Gonzalo Romero. Their valuable and constructive comments and feedbacks helped me improve the work in this thesis. Their advice and suggestions helped me grow as an academic.

I am fortunate to have my fellow Ph.D. students – Adam, Hossein, Jeff, Lu, Michael, Mohammad, Mojtaba, Vahid, Vahideh, and Yun – in this journey. They share with me the most joyful time in my life.

Finally, I am very grateful to my parents for their unconditional love and support. I thank them for always believing in me.

Contents

Chapter 1

Introduction

In this thesis, I study three questions in operations management. Specifically, the second chapter discusses the reorder flexibility in supply chain management and the last two chapters focus on operation strategies in service industry.

In Chapter 2, I examine the value of reorder flexibility under price competition. The supply chain literature shows that reorder flexibility increases profits under competition, assuming fixed prices or quantity competition. I show that price competition, arguably a more appropriate price formation model in the presence of reorder flexibility, may yield *opposite* results. I consider a three-stage model of duopoly firms that sell differentiated products with stochastic demand. Firms make reorder-flexibility decisions and then place initial orders, before observing demand. After observing demand, firms set prices and, if they have the option, may reorder at a higher cost. I show that the expected profit functions are not unimodal and provide extensive equilibrium results. These appear to be the first for stochastic finite-horizon price-inventory competition with more than one order opportunity. I show: (i) Unilateral reorder flexibility is *not* an equilibrium. (ii) Reorder flexibility may *increase* initial orders. (iii) Reorder flexibility *hurts profits except* if it reduces initial orders and in addition, demand variability is moderate, reordering is sufficiently inexpensive, and products are sufficiently differentiated. (iv) Firms can avoid the downside of reorder flexibility only in some cases where it hurts profits. In others, firms are trapped in a prisoner's dilemma, whereby *reorder flexibility is the dominant strategy even though it hurts their profits.*

In Chapter 3, I study how the growing prevalence of real-time delay information has an impact on a service system. I consider a single-server queueing system where customers arrive according to a Poisson process and service takes an exponential time. There are two streams of customers, one informed about real-time delay and one uninformed. I characterize the equilibrium behavior of customers who may balk in such a system and

investigate how a larger fraction of informed customers affects the system performance measures, i.e., throughput and social welfare. I show that the impacts of growing information prevalence on system performance measures are determined by the equilibrium joining behavior of *uninformed* customers. Perhaps surprisingly, we find that throughput and social welfare can be *unimodal* in the fraction of informed customers. In other words, some amount of information heterogeneity in the population can lead to strictly more efficient outcomes, in terms of the system throughput or social welfare, than information homogeneity. For example, under a very mild condition, throughput of a system with offered load being 1 will always suffer if there are more than 58% of informed customers in the population. Moreover, it is shown that for an overloaded system with offered load sufficiently higher than 1, social welfare always reaches its maximum when some fraction of customers are uninformed of the congestion in real time.

In the last chapter, I consider capacity allocation under endogenous arrivals of strategic customers with heterogeneous time-of-service (TOS) preferences. Literature on the design and control of congestion-prone service systems usually assumes that the customer arrival processes are exogenous, and in particular, independent of the time-varying wait time conditions. However, customers may account for system performance in choosing their arrival times. That is, they may adjust their visit times in exchange of shorter delays. In this paper, I propose a discrete time choice model that captures how rational customers with heterogeneous TOS preferences and delay sensitivities determine their arrival times. I show the existence of a customer choice equilibrium. Taking into account this equilibrium, I then consider the optimal intertemporal capacity allocation decision. I characterize the properties of the socially optimal solution and discuss its incentive compatibility with customer self-interested TOS choices. I find that with a limited total capacity, the optimal capacity allocation alone is sufficient to prevent the system efficiency loss from customer decentralized decisions. However, when the capacity costs play a role, the provider has to impose a pricing scheme to align customer incentives. Specifically, if the capacity costs are time-varying, the provider has to charge, in order to retain the incentive compatibility, for TOS such that price differences across time periods equal the corresponding capacity cost differences.

Chapter 2

The Peril of Reorder Flexibility under Price Competition

2.1 Introduction

Fashion apparel retailers face the problem of matching supply with demand over a short and unpredictable selling season. This is particularly challenging if they must make procurement decisions well before the season. However, the accuracy of demand forecasts increases dramatically at the inception of the season, when fashion trends are better understood. Retailers can exploit this improved demand information through *price flexibility* and *reorder flexibility*.

Price flexibility allows retailers to adapt prices to market conditions, contingent on whether demand for an item is high or low. Such flexibility also supports downside volume flexibility, whereby a retailer charges more than the liquidation price, to sell only a fraction of its inventory and hold back the rest. This strategy may boost profits under low demand. Some retailers are unwilling to slash prices even if this means throwing away unsold garments (Dwyer 2010).

However, price flexibility only helps manage demand. To adapt their *supply* and stay on top of fashion trends, reorder flexibility, a core component of *quick response* capabilities, is critical for fashion retailers. Their supply chains have become nimbler than ever. As a result of reduced lead times, they can order not only well in advance, but also place and receive additional (typically more expensive) orders right before or at the inception of the season, once better demand information is available. This upside volume flexibility also offers *downside protection*, as it allows firms to reduce initial orders and still maintain the ability to satisfy higher demand. For example, in the well-known Sport

Obermeyer case (Fisher and Raman 1996), the ability to place a second order reduces overstocking and increases product availability. With their lightning-fast supply chains, fast fashion retailers such as Zara are pushing reorder flexibility to an extreme.

Virtually the entire academic and trade literature focuses on these *benefits* of price and/or reorder flexibility for better matching supply with demand. However, this perspective ignores the following *downside* that we study in this paper: Reorder flexibility may erode profits by fostering more aggressive price competition. This effect is consistent with estimates of management consultants A.T. Kearney, that nowadays apparel retailers sell between 40 and 45 percent of inventory at a promotional price, up from 15 to 20 percent a decade ago (D'Innocenzio 2012). This paper cautions that, given the prevalence of price flexibility in the industry, concerns over intensified price competition are likely to grow as reorder flexibility proliferates. We address two main questions: (1) Why and under what conditions does reorder flexibility hurt or increase profits under price competition? (2) Which reorder flexibility configuration do retailers choose in equilibrium?

We study these questions in the context of a three-stage duopoly model with stochastic demand. Each firm sells a single differentiated product. Demand is a linear function of prices with an ex ante unknown intercept. Both firms simultaneously make reorder-flexibility decisions and then simultaneously place initial orders, before demand uncertainty resolves. In the last stage, after observing demand, firms simultaneously set prices and – if they have the option – reorder more units. The reorder unit cost exceeds the unit cost on initial orders, as shorter lead times increase sourcing and distribution costs. Leftover inventory is disposed with zero salvage value.

2.1.1 Overview of Main Results

This paper contributes novel managerial insights and technical results that derive from its focus on *price competition* under both *downside and upside volume flexibility*. Managerially, we identify under what conditions reorder flexibility hurts profits, and the resulting equilibrium flexibility choices. Technically, to our knowledge this paper provides the *first* equilibrium results for a stochastic finite-horizon problem under price competition with more than one order opportunity. In contrast, prior flexibility studies assume *quantity competition*. This distinction is important. Managerially, our results are in sharp contrast to those under quantity competition, and price competition may be a more appropriate model of price formation when firms have volume flexibility. Technically, our analysis overcomes challenges that arise only under price competition.

1. Reorder flexibility can hurt profits under price competition. The ordering and pricing equilibria under the symmetric reorder flexibility configurations yield the following results (see Sections 2.3-2.5).

(a) Orders. The key effect of price competition is that it leads to more aggressive ordering. This weakens or even reverses the downside protection of reorder flexibility. Specifically, only if reordering is sufficiently cheap, relative to early procurement, do firms with reorder flexibility order less initially than inflexible firms. Otherwise, however, flexible firms order *more initially* than inflexible firms. Furthermore, even in cases where flexible firms order less initially, if demand turns out high, they reorder so much that they end up with more inventory than inflexible firms.

(b) Expected profits. Reorder flexibility hurts profits whenever it yields larger initial inventories and therefore lower prices and higher procurement costs. However, reorder flexibility may also hurt profits if it leads to smaller initial orders. Two countervailing effects are at work under low versus high demand. Reorder flexibility hurts profits *except* if the gains from downside protection under low demand dominate the losses from intensified competition under high demand, which holds under three conditions: (*i*) products are sufficiently differentiated; (*ii*) the demand variability is neither too small nor too large; and (*iii*) reordering is sufficiently cheap.

2. Reorder flexibility configurations in equilibrium. We show that in the flexibility-selection stage that precedes the procurement-pricing decisions, unilateral reorder flexibility is *not* an equilibrium. Furthermore, bilateral inflexibility is the Pareto-dominant symmetric equilibrium only in some of the cases where reorder flexibility hurts profits. In these cases the firms can avoid the downside of reorder flexibility by committing to inflexibility. However, in other cases they are trapped in a prisoner's dilemma, whereby it is the *dominant strategy for firms to select reorder flexibility even though it hurts their profits* (see Section 2.6).

These results point to the strategic importance of product differentiation and efficient reorder operations as *complementary* capabilities, not only to reap the benefits of reorder flexibility, but also to avoid its downside.

3. Price competition versus quantity competition. Our results under price competition are in stark contrast to prior findings on the effects of volume flexibility under *quantity (Cournot) competition.*

(a) Upside volume flexibility can hurt profits only under price competition. Lin and Parlaktürk (2012) consider a reorder option for duopoly retailers that sell a homogeneous product under quantity competition. In their analysis competition between "fast" retailers that can reorder after demand is realized yields (weakly) smaller initial orders

and larger expected profits, compared to competition between "slow" retailers without a reorder option. These results are the *opposite* of our findings that the "reorder" game may yield larger initial orders and lower expected profits than the "no reorder" game, and moreover, that a reorder option can only benefit firms if products are sufficiently differentiated (see Sections 2.5.2-2.5.3).

(b) Downside volume flexibility may yield lower inventories and profits under price versus quantity competition. Anupindi and Jiang (2008) prove that competition among homogeneous-product firms with downside volume flexibility through a hold back option yields higher expected profits than competition among inflexible firms that sell all supply ex post at the clearance price (Van Mieghem and Dada 1999 show this numerically). In their model of flexible firms, equilibrium sales quantities and prices are determined under quantity competition[1]. We show that compared to quantity competition, under price competition flexible firms not only get lower profits, as expected, but they may also make *lower* inventory investments, because price competition reduces their control of downside risk through the hold back option (see Section 2.3.2).

The assumption of quantity competition is usually justified with the classic result that single-stage quantity competition yields the same outcome as two-stage competition where firms first choose supply quantities (capacity, production or inventory) and then prices (Kreps and Scheinkman 1983). However, this equivalence critically hinges on the condition that firms *cannot* increase their supply while or after demand is formulated (cf. Tirole 1998, p. 217). By its very nature, volume flexibility may clearly violate this condition, in which case price competition may be a more appropriate model of price formation. This would certainly seem to apply to firms such as Zara that can flexibly increase their inventories above initial levels after observing demand. Whether quantity or price competition is more appropriate under volume flexibility depends on factors that affect how flexibly firms can increase their supply, such as the marginal cost and the delivery time of replenishment orders (cf. Tirole 1998, p. 224). The contrast between the results under price versus quantity competition underscores the importance of understanding these factors in order to better predict and improve performance under volume flexibility.

4. Equilibrium analysis under price competition. We provide extensive results that explicitly characterize the equilibria in terms of the demand and cost characteristics. The analysis is challenging because under price competition, each firm's first-stage expected profit function is generally *not unimodal* in its own order. Quantity competition is much

[1] They show that quantity competition has the same outcome as a two-stage production-price subgame, which proves the stochastic counterpart of the result of Kreps and Scheinkman (1983).

more tractable as it does not face this challenge (see Section 2.3.1). This may explain why our results appear to be the first for stochastic finite-horizon price competition with more than one order. Our differentiated-products model has the dual appeal of being more plausible and more tractable under price competition than the homogeneous-product model that is prevalent under quantity competition.

2.1.2 Literature Review

This paper is at the intersection of two literatures. One studies stochastic price-inventory control problems, the other considers the impact of price and/or operational flexibility on profitability.

Numerous papers in both streams focus on *monopoly* settings. See Chen and Simchi-levi (2012) for a recent survey of integrated price-inventory control models and Petruzzi and Dada (2011) who focus on newsvendor models. Among monopoly studies of flexibility, Van Mieghem and Dada (1999) study the benefits of production and price postponement strategies, with limited analysis of quantity competition; Cachon and Swinney (2009) show that quick response can be significantly more valuable to a retailer in the presence of strategic consumers than without them; Goyal and Netessine (2011) analyze volume and product flexibility under endogenous pricing.

The understanding of *stochastic price-inventory control under competition* is limited. At one extreme of the problem space, Bernstein and Federgruen (2005) and Zhao and Atkins (2008) study the *single period* problem in the classic newsvendor framework: The selling period is so short compared to lead times that firms can order only once, before demand is realized, and also choose prices in advance. At the other extreme, Kirman and Sobel (1974) and Bernstein and Federgruen (2004) study periodic-review *infinite-horizon* oligopolies, but under conditions that reduce them to myopic single period problems where decisions in each period are made before demand is realized. Kirman and Sobel (1974) obtain a partial characterization of a pure strategy Nash equilibrium with a stationary base-stock level. Bernstein and Federgruen (2004) identify conditions for existence of a pure strategy Nash equilibrium in which each retailer adopts a stationary base-stock policy and list price. Our paper studies models in the intermediate domain between the single-period and infinite-horizon extremes: The selling horizon is *finite* but firms are sufficiently responsive to exploit information gained over time to *make decisions more than once*. This case is gaining importance as businesses are countering shrinking product lifecycles with faster operations and adaptive pricing. However, this appears to be the *first* paper that considers price competition among firms that can *order more than*

once over time. In the stochastic Bertrand-Edgeworth model (cf. Hviid 1991, Reynolds and Wilson 2000) homogeneous-product firms *order only once*, before observing demand, and set prices thereafter. In other studies (cf. Porteus et al. 2010, Liu and Zhang 2013 and references therein) firms only compete in prices whereas capacity levels are *exogenous*.

In the literature on *flexibility under competition*, a number of economics papers study the strategic effects of flexibility in the absence of demand uncertainty, e.g., Maggi (1996), Boccard and Wauthy (2000), Röller and Tombak (1993). A general insight from these studies is that flexibility can be harmful under competition. However, demand uncertainty is the key challenge to matching supply with demand, and the fundamental reason for supply chain flexibility. Studies of flexibility under competition with stochastic demand can be grouped into two streams. One focuses on product flexibility (Anand and Girotra 2007, Goyal and Netessine 2007), the other, which includes this paper, on volume flexibility (cf. Vives 1989, Van Mieghem and Dada 1999, Anupindi and Jiang 2008, Li and Ha 2008, Caro and Martínez-de-Albéniz 2010, Lin and Parlaktürk 2012). In contrast to this paper, none of these studies consider price competition: They either assume *fixed* prices or consider endogenous pricing under assumptions that lead to *quantity* competition.

In the product flexibility stream, Anand and Girotra (2007) consider two-product firms that choose between early product differentiation before, or delayed differentiation after demand uncertainty is resolved. They show that early differentiation may arise as a dominant strategy. Goyal and Netessine (2007) consider two-product firms that choose whether to invest in flexible or dedicated technology and identify conditions under which flexibility benefits or harms profits. In both papers, unlike in ours, prices are determined by quantity competition and the total supply is determined before demand uncertainty resolves; the essence of product flexibility is that it allows firms to delay product-to-market allocation decisions until demand is known.

The volume flexibility stream has more of a history in economics. Going back to Stigler (1939), these papers often model the degree of flexibility on a continuum, by the slope of the average cost curve around some minimum; cf. Vives (1989) who studies a two-stage homogeneous-product oligopoly in which firms choose their flexibility level before receiving (private) demand signals, and then choose production quantities. In contrast, volume flexibility studies in the operations management literature typically consider two discrete flexibility configurations that differ in terms of the timing of supply decisions (capacity/production/inventory) relative to when demand is realized. The key finding that is common to these papers is that competition with volume flexibility *increases* expected profits compared to competition without such flexibility: This holds for downside

flexibility through a hold back option under quantity competition (Van Mieghem and Dada 1999, Anupindi and Jiang 2008), and for upside flexibility through a reorder option or reactive capacity – both under fixed prices (Li and Ha 2008, Caro and Martínez-de-Albéniz 2010) and under quantity competition (Lin and Parlaktürk 2012). As discussed in Section 2.1.1, the conditions for the equivalence of price and quantity competition (Kreps and Scheinkman 1983, Anupindi and Jiang 2008) may not hold in the presence of volume flexibility. We show that the results under price competition may be the *opposite* of those under quantity competition and fixed prices.

Wu and Zhang (2014) study a sourcing game where homogeneous-product firms first choose between efficient (long lead-time, low cost) and responsive (short lead-time, high cost) sourcing, then place orders, and finally, after demand is realized, sell their inventories at the market-clearing price. Their setup bears some resemblance to ours, but there are important differences. In terms of modeling, we study price competition, and more importantly, we allow *two* orders, early and late, whereas in their model firms can order only *once*, early or late. Our model may be more applicable for fashion retailers such as Sports Obermeyer or Zara that typically order more than once. The results of the two papers are therefore not directly comparable, and they focus on different issues. Indeed, in their model, even without competition, either sourcing option may be preferred depending on the cost-information tradeoff. Wu and Zhang (2014) study the effects of competition and information on this tradeoff. In contrast, in our model a monopoly always prefers reorder flexibility, and we identify under what conditions price competition reverses this preference.

2.2 Models, Problem Formulations, and Preliminary Analysis

We study duopoly firms with price flexibility, each selling a single differentiated product with price-sensitive demand. The bulk of the paper (Sections 2.2-2.5) focuses on the analysis and comparison of two games in which the reorder option is symmetric between firms. In the "no reorder" game, referred to as N game, firms have no reorder flexibility but *only* price flexibility. In the "reorder" game, referred to as R game, firms have price *and* reorder flexibility. In Section 2.6 we justify this focus on symmetric flexibility configurations: We show that unilateral reorder flexibility is *not* an equilibrium in the flexibility-selection stage that precedes the procurement-pricing decisions.

2.2.1 Models

The N and R games share the following two-stage structure. In stage one, before observing demand, firms simultaneously choose their initial orders. The outcome of stage one is common knowledge. Demand uncertainty is resolved before firms make their second-stage decisions. The N and R games are identical up to this instant when demand is observed. They differ as follows in the second-stage decisions. In the N game, firms simultaneously choose prices but they cannot reorder. In the R game, firms simultaneously choose prices *and* reorder quantities. The initial unit procurement cost is typically lower than the reorder unit cost. Demand and sales occur following the second-stage decisions. Taken literally, this captures a situation where firms gain demand information through factors other than their own early-season sales, such as weather, market news and fashion trends. However, the model can also be viewed as a reasonable approximation of settings where sales that materialize between the first order delivery and the second-stage decisions only make up a small fraction of initial inventory but are still of significant value for demand forecasting. It is quite common that the forecast accuracy for total season demand increases dramatically after observing a few days of early season sales. The model does not specify delivery lead times; we assume they are short enough so firms do not lose sales due to delivery delays. Without loss of generality the salvage value of leftover inventory is zero. We ignore further holding costs that may be incurred during the season, as they are insignificant relative to margins and overstocking costs.

We index the firms by $i \in \{1, 2\}$ and denote their variables and functions, such as order quantities, prices, demands, and profits, by corresponding subscripts. We write $-i$ to denote firm i's rival, where $-i \neq i$. We model demand uncertainty in a linear demand system, which is widely used in the literature on differentiated products (e.g., McGuire and Staelin 1983, Singh and Vives 1984). The linear form arises as the solution of the optimal consumption problem of a representative consumer with quadratic utility (Vives 2001, Chapter 6.1). Let p_i denote firm i's price and $\mathbf{p} = (p_1, p_2)$. Demand for firm i's product is given by

$$d_i(\mathbf{p};\alpha) = \alpha - p_i + \gamma p_{-i} \geq 0, \ i = 1, 2. \tag{2.1}$$

The intercept $\alpha > 0$ is ex ante uncertain; we call it the *market size parameter*. We assume that demand is high, i.e., $\alpha = \alpha_H$, or low, i.e., $\alpha = \alpha_L < \alpha_H$, with equal probability. The assumption of equally likely high- and low-demand scenarios does not change our main qualitative insights. Uncertainty in α may be due to factors that equally affect differentiated products in the same category, such as color in the case of fashion items. The *product substitution parameter* $\gamma \in [0, 1)$ reflects the degree of

product differentiation and factors such as brand preferences. We assume that γ is known, based on the notion that brand loyalty and price sensitivity are well understood. Firm-i's demand sensitivity to its rival's price increases in γ. The larger γ the less differentiated the products. For $\gamma \approx 1$ each firm's demand is (approximately) equally sensitive to both prices. The demand system (2.1) does not explicitly model situations with *perfectly* substitutable products (the higher-price firm gets positive demand even as $\gamma \to 1$), but as discussed in Section 2.7.1, our main insights continue to hold for a scaled version of (2.1) that does capture perfect substitution. At last, the parsimonious demand model in (2.1) represents the dependence of demand on price in a scaled system as that in McGuire and Staelin (1983). It is thus invalid to directly compare the actual prices, quantities, and profits across different values of γ in the demand system defined in (2.1). For example, one may find that the demand model (2.1) implies that the aggregate (industry) demand increases with γ, which contradicts with our intuition that the amount of potential customers who are interested in either product at least does not increase as the two products are less differentiated. In fact, this is true for our scaled model (2.1) after mapping it back to the real quantities. We refer to Staelin (2008) for more detailed discussions. Note that we establish all our results based on the assumption that γ is a predetermined parameter. As a result, the parsimonious scaled demand system (2.1) does not alter our results.

In stage one, before knowing whether the market size will be α_L or α_H, firm i chooses its (initial) order x_i at unit cost $c \in [0, C]$, where C is the unit reorder cost that is available in the presence of reorder flexibility. We normalize $C \equiv 1$ in our analysis without loss of generality, but use the notation C in our discussion. Let $\mathbf{x} = (x_1, x_2)$ denote the initial order vector.

2.2.2 Problem Formulations

Both for the N and the R game, our analysis focuses on pure-strategy Nash equilibria in the second stage and on subgame-perfect Nash equilibria in symmetric order strategies in the first stage.

No reorder game. Let $\pi_i^N(\mathbf{p}, x_i; \alpha)$ denote firm i's second-stage revenue function in the N game. It depends on both prices $\mathbf{p}$, firm i's initial inventory x_i, and the realized market size α. Given initial inventories $\mathbf{x}$ and market size α, firms simultaneously choose prices in the second stage:

$$\max_{p_i} \ \pi_i^N(\mathbf{p}, x_i; \alpha) = p_i \cdot \min\left(x_i, d_i(\mathbf{p};\alpha)\right), \ i = 1, 2, \tag{2.2}$$

where $\pi_i^N(\mathbf{p}, x_i; \alpha)$ is strictly concave in p_i. We assume that excess demand is lost. Since firms observe the market size realization α prior to choosing prices, they have no incentive to generate more demand than they can satisfy. However, a firm may find it optimal not to sell all the inventory procured in the first stage, if the market turns out to be small. In this case, we say that the firm prices to *hold back* inventory. As noted above, leftover inventory has zero salvage value. Let $\mathbf{p}^{N*}(\mathbf{x};\alpha)$ denote the second-stage equilibrium price vector and $\pi_i^{N*}(\mathbf{x};\alpha) = \pi_i^N\left(\mathbf{p}^{N*}(\mathbf{x},\alpha), x_i;\alpha\right)$ firm i's second-stage equilibrium revenue for the N game, as a function of the initial orders $\mathbf{x}$ and the market size α. Let $\Pi_i^N(\mathbf{x})$ denote firm i's expected profit as a function of initial orders $\mathbf{x}$. In the first stage, firms simultaneously choose their orders by solving

$$\max_{x_i \geq 0} \ \Pi_i^N(\mathbf{x}) = \frac{1}{2}\left(\pi_i^{N*}(\mathbf{x};\alpha_L) + \pi_i^{N*}(\mathbf{x};\alpha_H)\right) - cx_i, \ i = 1, 2. \tag{2.3}$$

Let $\mathbf{x}^{N*}$ denote equilibrium orders, and the scalar x^{N*} a symmetric equilibrium order quantity.

Reorder game. Let $\pi_i^R(\mathbf{p}, x_i;\alpha)$ denote firm i's second-stage profit function in the R game. In addition to choosing prices, firms can reorder inventory at a unit cost $C \geq c$. The reorder quantities are determined by the initial inventories and the prices: given x_i and $\mathbf{p}$, firm i orders the amount $(d_i(\mathbf{p};\alpha) - x_i)^+ = \max(d_i(\mathbf{p};\alpha) - x_i, 0)$ in the second stage. Given initial inventories $\mathbf{x}$ and market size α, firms simultaneously choose prices and the resulting reorder quantities in the second stage:

$$\max_{p_i} \ \pi_i^R(\mathbf{p}, x_i;\alpha) = p_i d_i(\mathbf{p};\alpha) - C(d_i(\mathbf{p};\alpha) - x_i)^+, \ i = 1, 2, \tag{2.4}$$

where $\pi_i^R(\mathbf{p}, x_i;\alpha)$ is concave in p_i. Let $\mathbf{p}^{R*}(\mathbf{x};\alpha)$ denote the second-stage equilibrium price vector and $\pi_i^{R*}(\mathbf{x};\alpha) = \pi_i^R\left(\mathbf{p}^{R*}(\mathbf{x};\alpha), x_i;\alpha\right)$ firm i's second-stage equilibrium profit function for the R game. Let $\Pi_i^R(\mathbf{x})$ denote firm i's expected profit for the R game as a function of the initial order vector. In the first stage, firms simultaneously choose their initial orders by solving

$$\max_{x_i \geq 0} \ \Pi_i^R(\mathbf{x}) = \frac{1}{2}\left(\pi_i^{R*}(\mathbf{x};\alpha_L) + \pi_i^{R*}(\mathbf{x};\alpha_H)\right) - cx_i, \ i = 1, 2. \tag{2.5}$$

Let $\mathbf{x}^{R*}$ denote equilibrium initial orders, and the scalar x^{R*} a symmetric equilibrium initial order.

In both games, each firm's expected profit may be bimodal in its own initial order, due to the *joint* effect of price competition, uncertainty, and the sequential decisions

over a finite horizon (see Section 2.3.1). Therefore, our analysis cannot rely on standard equilibrium characterization results. To overcome this challenge we exploit the structure of the best response problem.

Terminology. We refer to the option to hold back inventory also as *downside volume flexibility*. Firms have this flexibility in both games. We refer to the option to reorder inventory also as *upside volume flexibility*. Firms have this flexibility only in the R game. Compared to a hold back option, a reorder option offers firms stronger downside protection against low demand: Firms that hold back a fraction of initial orders still incur the cost on all ordered units. In contrast, the reorder option allows firms to *reduce* initial orders (and costs) and still maintain the ability to satisfy higher demand. Therefore, we use the phrase *downward protection* only in reference to the latter scenario.

2.2.3 Second-Stage Price-Reorder Equilibria

As a preliminary analysis we characterize the second-stage subgame equilibria of the N and R games. These serve as building blocks for our characterization and comparison of the first-stage order equilibria in Sections 2.3-2.6. Since firms learn the market size α prior to their second-stage decisions, the second-stage subgames are deterministic. Each subgame has an unique equilibrium that depends as follows on the initial order vector $\mathbf{x}$ and the realized market size α (Lemma 2.1 below summarizes these results).

No reorder game. Given initial inventories $\mathbf{x}$ and the realized market size α, firms simultaneously choose prices by solving (2.2). Define firm i's *clearance price* $p_i^c(x_i, p_{-i}; \alpha)$ as the highest price that generates enough demand to sell its inventory x_i, given its rival charges p_{-i} and the market size is α. Firm i may choose to charge more than this clearance price and hold back supply. Define firm i's *hold back price* $p_i^h(p_{-i}; \alpha)$ as its revenue-maximizing price in the absence of inventory constraints. Firm i prefers this price if it exceeds the clearance price, which results in leftover stock. That is, firm i's best response price is the larger of its clearance and hold back prices. We call these best responses *clearance (c)* and *hold back (h)*, respectively.

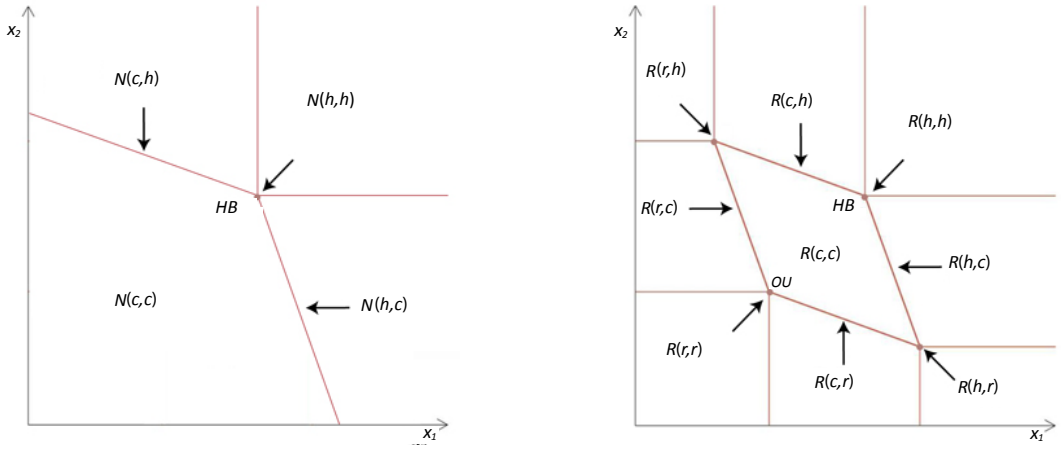
Both firms follow these strategies. As shown in Figure 2.1a the firms' equilibrium strategies partition the initial inventory space $\{\mathbf{x} \geq \mathbf{0}\}$ into four regions. The labels $N(c,c)$, $N(h,c)$, $N(c,h)$ and $N(h,h)$ identify the equilibrium strategies for each region, the first letter refers to firm 1 and the second to firm 2. For example, for initial inventory vectors $\mathbf{x} \in N(c,c)$ the unique equilibrium is for each firm to charge its clearance price. For high initial inventories $\mathbf{x} \in N(h,h)$ the unique equilibrium is for each firm to charge the hold back price that generates demand and sales equal to the hold back threshold

$HB := \alpha/(2-\gamma)$, so each firm has leftover inventory. The arrows in 2.1a indicate the shift from an initial inventory vector to the corresponding equilibrium sales.

Reorder game. Given initial inventories $\mathbf{x}$ and the realized market size α, firms simultaneously choose their prices and reorder quantities by solving (2.4). *Clearance* and *hold back* are two possible best responses, as in the N game. A third possibility is for firm i to charge the *reorder price* $p_i^r(p_{-i};\alpha)$, defined as its profit-maximizing price if it has no initial inventory. (The reorder price exceeds the hold back price since reordering is costly.) If its reorder price is lower than its clearance price, firm i's best response is to charge the reorder price and to procure more units to satisfy its excess demand at that price; we call this the *reorder (r)* strategy. Otherwise, firm i's best response is not to reorder and to charge the larger of its clearance and hold back prices.

Both firms follow these strategies. The resulting equilibrium strategies partition the initial inventory space $\{\mathbf{x} \geq \mathbf{0}\}$ into nine regions as shown in Figure 2.1b. We use the same labeling convention as in the N game. For example, for initial inventories $\mathbf{x} \in R(r,r)$, the unique equilibrium is for each firm to charge the reorder price that generates demand equal to the order-up-to level $OU := (\alpha - C(1-\gamma))/(2-\gamma)$, and to reorder up to and sell this amount, as indicated by the arrow. For $\alpha \leq C(1-\gamma)$ firms have no incentive to reorder regardless of their initial inventory levels, and the second-stage N and R subgames are equivalent.

Figure 2.1: Second Stage Equilibrium Strategies as Functions of Initial Inventories



(a) No Reorder Game (N) (b) Reorder Game (R)

Lemma 2.1 summarizes these results (see proof for closed-form prices and quantities).[2]

[2] These N and R games are also analyzed in Maggi (1996); however, he restricts attention to deterministic demand. His second stage equilibrium characterization is less complete and explicit than what we require in Sections 2.3-2.6 for our analysis under stochastic demand. Therefore, we provide in Lemma

Lemma 2.1 (Second Stage Subgame Equilibria). *For any initial inventory vector* $\mathbf{x}$ *and market size realization* α*, the price subgame of the* N *game and the price-reorder subgame of the* R *game each has a unique equilibrium. The equilibrium strategy of firm* i *depends as follows on* $\mathbf{x}$*.*

(N) In the N *game there is a hold back threshold* $\overline{x}_i(x_{-i};\alpha)$ *such that: (i) if* $x_i \leq \overline{x}_i(x_{-i};\alpha)$ *then firm* i *prices to clear its inventory; (ii) if* $x_i > \overline{x}_i(x_{-i};\alpha)$ *then it prices to hold back inventory and sells* $\overline{x}_i(x_{-i};\alpha)$*.*

(R) In the R *game there are hold back and order-up-to thresholds* $\underline{x}_i(x_{-i};\alpha) < \overline{x}_i(x_{-i};\alpha)$ *such that: (i) if* $x_i < \underline{x}_i(x_{-i};\alpha)$ *then firm* i *reorders up to* $\underline{x}_i(x_{-i};\alpha)$ *and charges the reorder price to sell this amount; (ii) if* $\underline{x}_i(x_{-i};\alpha) \leq x_i \leq \overline{x}_i(x_{-i};\alpha)$ *then it prices to clear its inventory but does not reorder; (iii) if* $x_i > \overline{x}_i(x_{-i};\alpha)$ *then it prices to hold back inventory and sells* $\overline{x}_i(x_{-i};\alpha)$*.*

In Sections 2.3 and 2.4 we characterize the first-stage order equilibria for the N and R games, respectively. These results build on Lemma 2.1.

2.3 Price Flexibility without Reorder Flexibility

In this section we first characterize the N game equilibria and then compare these results with those under quantity competition.

2.3.1 Downside Volume Flexibility under Price Competition: The N Game

In the first stage, firms simultaneously choose their order quantities by solving (2.3), where the second-stage equilibrium revenue functions $\pi_i^{N*}(\mathbf{x};\alpha_L)$ and $\pi_i^{N*}(\mathbf{x};\alpha_H)$ are given by Lemma 2.1.

Price competition, uncertainty, and sequential decisions imply bimodal expected profits. The equilibrium characterization of the N game is significantly complicated by the fact that the second-stage equilibrium revenue functions $\pi_i^{N*}(\mathbf{x};\alpha_L)$ and $\pi_i^{N*}(\mathbf{x};\alpha_H)$ are not concave in firm i's own order quantity. This fact, combined with demand uncertainty, implies that the first-stage expected profit functions of each firm may be bimodal in its own order.

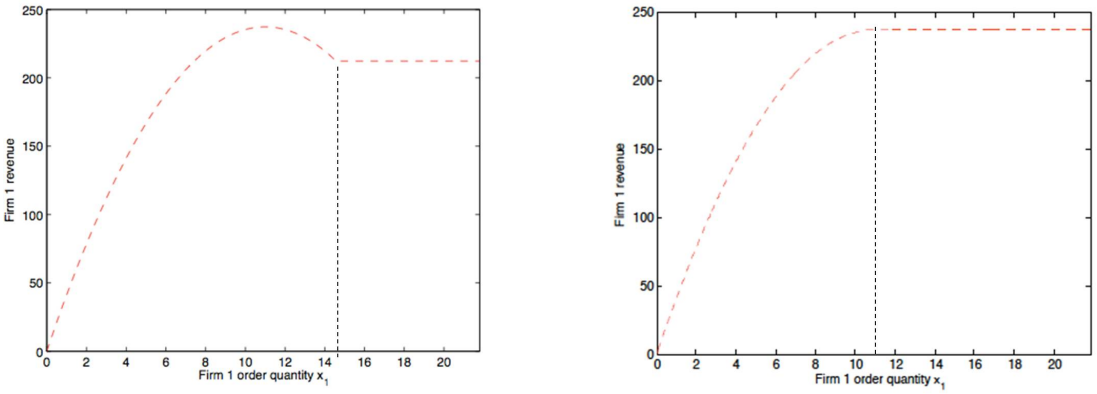
The non-concave nature of the second-stage equilibrium revenue functions is the natural result of *price competition*, coupled with the sequential nature of decisions over a

2.1 our own results and proof.

finite horizon. Consider how $\pi_i^{N*}(\mathbf{x};\alpha)$ depends on firm i's order x_i, given the market size is α and its rival orders x_{-i}. By Lemma 2.1, in the N game equilibrium firm i never sells more than $\overline{x}_i(x_{-i};\alpha)$; it holds back any inventory in excess of this threshold. The function $\pi_i^{N*}(\mathbf{x};\alpha)$ is concave in $x_i \leq \overline{x}_i(x_{-i};\alpha)$, constant in $x_i \geq \overline{x}_i(x_{-i};\alpha)$, and peaks at a *smaller* quantity than the hold back threshold $\overline{x}_i(x_{-i};\alpha)$. Figure 2.2a shows a representative example for firm 1, given $\alpha = 15$ and $x_2 = 5$. Firm 1's second-stage equilibrium revenue peaks at $x_1 = 11.0$, then decreases, before leveling off at the hold back threshold $\overline{x}_1(x_2;\alpha) = 14.6$. The property that the hold back threshold exceeds the revenue-maximizing quantity is due to price competition: The marginal revenue of each firm is higher if it unilaterally drops its price than if it unilaterally increases its inventory, so equilibrium prices keep dropping as x_1 increases from the revenue-maximizing quantity $x_1 = 11.0$ to $x_1 = 14.6$. Therefore, firm 1 has an incentive to sell more than 11 units if it has the inventory, even though doing so hurts its revenue.

In contrast, if prices are determined by *quantity competition*, then each firm's second-stage equilibrium revenue function is *concave* in its own order quantity and peaks at the hold back threshold, unlike under price competition. Figure 2.2b shows a representative example. Given $\alpha = 15$ and $x_2 = 5$, firm 1's hold back threshold equals $x_1 = 11$ units under quantity competition, and its second-stage equilibrium revenue function peaks at this threshold.

Figure 2.2: Effect of Competition Mode on Second Stage Equilibrium Revenue Function ($x_2 = 5$, $\alpha = 15$, $\gamma = 0.7$)



(a) Price Competition

(b) Quantity Competition

In summary, the first-stage expected profit functions may be bimodal (and more generally, not unimodal under an arbitrary market size distribution), which is due to the *joint* effect of three factors, price competition, demand uncertainty and the sequential nature of ordering and pricing decisions over a finite horizon. If any one of these factors

is absent, the payoff functions (in each stage) are well-behaved. This seems to be the case virtually throughout the literature, including in papers where the equilibrium prices are determined by quantity competition (cf. Anupindi and Jiang 2008), and in studies on joint price-inventory competition in *stationary infinite-horizon* problems (cf. Kirman and Sobel 1974, Bernstein and Federgruen 2004). A noteworthy exception is the version of our N game with perfectly *homogeneous* products (cf. Hviid 1991, Reynolds and Wilson 2000). In this extreme case, demand functions are discontinuous in prices, first-stage payoffs are not well-behaved, and a pure strategy symmetric equilibrium in inventories need not exist if the extent of demand variation exceeds a threshold level (Reynolds and Wilson 2000). In contrast, our N game with differentiated products admits the following equilibrium result.

Proposition 2.2 (First Stage Order Equilibria: N Game). *Two thresholds on the market size ratio $r_\alpha := \frac{\alpha_H}{\alpha_L}$ determine the first stage order equilibria in the "no reorder" game:*

$$r_\alpha^{**} := m^{**}(\gamma) + 2c(1-\gamma)/\alpha_L > r_\alpha^{*} := m^{*}(\gamma) + 2c(1-\gamma)/\alpha_L \text{ for } \gamma > 0,$$

where $m^*(\gamma)$ and $m^{**}(\gamma)$ are explicit functions of γ and $m^{**}(\gamma) > m^*(\gamma) > 1$ for $\gamma > 0$. Let $HB(\alpha) := \alpha/(2-\gamma)$ denote the hold back threshold for market size α. If the market size ratio is:

(i) below the smaller threshold, i.e., $r_\alpha \le r_\alpha^*$, there is a unique symmetric order equilibrium:

$$x^{N*} = x_l^{N*} := \frac{(1+\gamma)(\alpha_H/2 + \alpha_L/2 - c(1-\gamma))}{2+\gamma} \le HB(\alpha_L), \tag{2.6}$$

and firms price to clear their inventory in both demand scenarios;

(ii) larger than the larger threshold, i.e., $r_\alpha \ge r_\alpha^{**}$, there is a unique symmetric order equilibrium:

$$HB(\alpha_L) < x^{N*} = x_h^{N*} := \frac{1+\gamma}{2+\gamma}(\alpha_H - 2c(1-\gamma)) < HB(\alpha_H), \tag{2.7}$$

firms price to sell $HB(\alpha_L)$ and hold back inventory if demand is low, and they price to clear their inventory if demand is high;

(iii) between the two thresholds, i.e., $r_\alpha^* < r_\alpha < r_\alpha^{**}$, there are exactly two symmetric equilibria, one as in (i), the other as in (ii). Moreover, x_l^{N*} Pareto-dominates x_h^{N*}.

Given any symmetric initial inventories in the second stage, firms sell at most the hold back threshold $HB(\alpha)$ corresponding to the realized market size α. The equilibrium order is therefore smaller than $HB(\alpha_H)$, but it may be smaller or larger than $HB(\alpha_L)$ due to market size uncertainty.

"Small" equilibrium: $x_l^{N*} < HB(\alpha_L)$. If the high-demand market is not too large, relative to the low-demand market ($r_\alpha < r_\alpha^{**}$), the Pareto-dominant equilibrium order is such that firms price to clear their inventory under either market size realization. The equilibrium order x_l^{N*} in (2.6) and the expected equilibrium price and profit of each firm are the same as under the corresponding *riskless* problem, i.e., if the market size were known and equal to the mean $(\alpha_H + \alpha_L)/2$. However, market size uncertainty leads to variability in equilibrium profits, leaving firms better off under high demand and worse off under low demand, compared to the riskless case.

"Large" equilibrium: $x_h^{N*} > HB(\alpha_L)$. If the high-demand market is relatively large ($r_\alpha > r_\alpha^{**}$), the firms order more in equilibrium than in the corresponding riskless problem. Their order x_h^{N*} is so large that they only price to clear their inventory if demand is high. If demand is low, they charge the hold back price, sell $HB(\alpha_L)$ and have leftover inventory, so that their revenues are independent of how much the firms order in excess of $HB(\alpha_L)$. The equilibrium quantity x_h^{N*} in (2.7) balances the marginal ordering cost with the incremental revenue under high demand.

2.3.2 Downside Volume Flexibility under Price versus Quantity Competition

Earlier studies of volume flexibility with stochastic demand consider endogenous pricing under *quantity* competition. This assumption is usually justified with the classic result that single-stage quantity competition yields the same outcome as two-stage competition where firms first choose supply quantities (capacity, production or inventory) and then prices (Kreps and Scheinkman 1983). However, this equivalence critically hinges on the condition that firms *cannot* increase their supply while or after demand is formulated (cf. Tirole 1998, p. 217). By its very nature, volume flexibility may clearly violate this condition, in which case price competition may be a more appropriate model of price formation. Our novel price competition results are relevant in this light. They also allow us to compare the effects of volume flexibility under price versus quantity competition. Here we focus on downside volume flexibility, in Sections 2.5.2-2.5.3 on reorder flexibility.

In the N game firms have downside volume flexibility through a hold back option. Anupindi and Jiang (2008) prove for perfectly *homogeneous* products that competition

under downside volume flexibility through a hold back option yields higher expected profits and capacity investments than competition among inflexible firms. In their model inflexible firms choose supply quantities in the first stage, before demand is known, and sell all supply in the second stage at the clearance price, i.e., they have no hold back option. Flexible firms also choose capacities ex ante, but sales quantities and prices are determined in the second stage under *quantity competition*. (They show the quantity competition subgame to be equivalent to a two-stage production-pricing subgame).

To study how the effects of a hold back option depend on the mode of competition, we compare the results of our N game (with hold back under price competition) to those of two variations, the N *game with clearance*, and the N *game under quantity competition* (with hold back). These variations correspond to the models in Anupindi and Jiang (2008) of inflexible and flexible firms, respectively (they consider significantly more general demand uncertainty). The analysis of these N game versions is much simpler than under price competition, and they each have a unique symmetric pure strategy equilibrium. We omit the details[3] and summarize our results informally:

1. *Price competition reduces but does not eliminate the value of downside volume flexibility. The N game under quantity competition yields (weakly) higher expected profits than the N game under price competition, and both yield higher profits than the N game with clearance.*

2. *Price competition yields lower inventory investments than quantity competition under moderate demand variability: There is an unique threshold $r_\alpha^Q (< r_\alpha^* < r_\alpha^{**})$, such that if $r_\alpha \in (r_\alpha^Q, r_\alpha^{**})$ then the Pareto-dominant equilibrium order is x_l^{N*} in the N game under price competition, and the equilibrium order is $x_h^{N*} > x_l^{N*}$ in the N game under quantity competition.*

That price competition may yield *lower* inventories than quantity competition, runs counter to the deterministic case. It follows because under quantity competition, firms have better control of downside risk – they can hold back more inventory. Hence they invest more upfront, sell more under high demand and less (by exercising their hold back option) under low demand.

The power of the hold back option rests on the flexible firms' ability to commit to underutilizing capacity if demand is low. The extent of this hold back commitment in turn depends on the mode of competition in the second stage, following the capacity investments. A model where equilibrium sales quantities and prices are determined in

[3]Proofs of these statements are available upon request.

the second stage under quantity competition, as in Anupindi and Jiang (2008), captures settings where firms are not sufficiently flexible to deviate from second-stage production quantities even if they have excess capacity. However, if firms can flexibly supply in the second stage any amount up to their initial capacity choice – due to preproduction or production postponement with highly flexible operations, then no quantity competition equilibrium with excess capacity is sustainable: Each firm has the incentive to lower its price and increase sales. Price competition is more appropriate in such cases as no firm has an incentive to lower its price and increase sales at the corresponding equilibrium, even if it has the capacity to do so. In this sense the hold back commitment is more credible under the price competition equilibrium.

2.4 Price and Reorder Flexibility

In this section we characterize the initial order equilibria for the R game, i.e., under price and reorder flexibility. Firms simultaneously choose their initial orders by solving (2.5). The second-stage equilibrium profit functions $\pi_i^{R*}(\mathbf{x};\alpha_L)$ and $\pi_i^{R*}(\mathbf{x};\alpha_H)$ are specified by Lemma 2.1. As in the N game, due to price competition, each firm's first-stage expected profit function may be bimodal in its initial order. The R game analysis is further complicated by the reorder option.

We henceforth assume that $\alpha_L = C(1-\gamma)$, which implies that it is not profitable to reorder if demand is low (i.e., $OU(\alpha_L) = 0$). This assumption seems reasonable in that firms typically procure enough early on to cover at least what they consider to be their base demand. Relaxing this assumption makes the analysis more cumbersome without generating additional insights, as confirmed through extensive numerical experiments with different values of α_L.

Proposition 2.3 (First Stage Order Equilibria: R Game). *Consider the "reorder" game with $\alpha_L = C(1-\gamma) < \alpha_H$. There exists a symmetric initial order equilibrium x^{R*}. Under the strictly Pareto-dominant symmetric equilibrium, in the second stage the firms do not reorder under low demand and they price to clear inventory under high demand; their price-reorder strategies depend as follows on the market size ratio $r_\alpha := \alpha_H/\alpha_L$ and the order cost ratio $r_c := c/C$:*

Market Size Ratio	*Pricing Strategy under Low Demand*	*Reorder Strategy under High Demand*
(i) $r_\alpha \le m^{**}(\gamma)$	*clear inventory*	*reorder to* $OU(\alpha_H)$ *iff* $r_c > \bar{r}_c(r_\alpha,\gamma)$
(ii) $m^{**}(\gamma) < r_\alpha < 2$	*clear inventory if* $r_c \ge \underline{r}_c(r_\alpha,\gamma)$; *sell* $HB(\alpha_L)$ *with leftover otherwise*	
(iii) $r_\alpha \ge 2$	*clear inventory if* $r_c \ge r_c(r_\alpha,\gamma)$; *sell* $HB(\alpha_L)$ *with leftover otherwise*	*reorder to* $OU(\alpha_H)$ *iff* $r_c > r_c(r_\alpha,\gamma)$

The thresholds $m^{**}(\gamma)$, $\underline{r}_c(r_\alpha,\gamma)$, $\overline{r}_c(r_\alpha,\gamma)$, *and* $r_c(r_\alpha,\gamma)$ *are explicit functions and* $\underline{r}_c(r_\alpha,\gamma) < \overline{r}_c(r_\alpha,\gamma)$. *The hold back threshold under low demand, and the order-up-to level under high demand are, respectively,*

$$HB(\alpha_L) = C\frac{1-\gamma}{2-\gamma} \text{ and } OU(\alpha_H) := \frac{\alpha_H - C(1-\gamma)}{2-\gamma}.$$

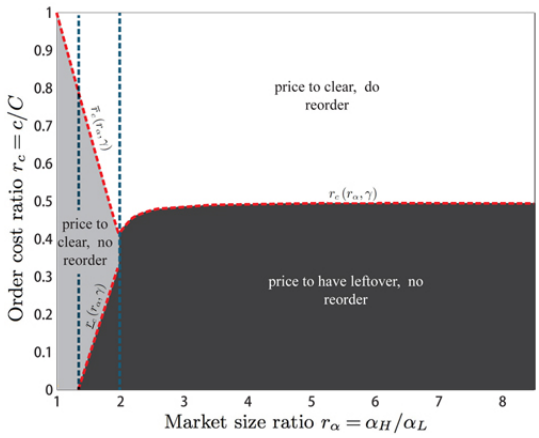
Proposition 2.3 identifies three combinations of low-demand pricing and high-demand reordering that can occur under a strictly Pareto-dominant equilibrium. The firms do not reorder under low demand since doing so is unprofitable for $\alpha_L = C(1-\gamma)$. They price to clear inventory under high demand because they have no incentive to order more than the high-demand hold back threshold $HB(\alpha_H)$. Note that, under no *strictly* Pareto-dominant equilibrium do firms in the second stage price to have leftover under low demand but reorder under high demand. These strategies hold only for initial orders in the range $(HB(\alpha_L), OU(\alpha_H))$. However, firms weakly prefer ordering initially at most $HB(\alpha_L)$ or at least $OU(\alpha_H)$, because their second-stage equilibrium revenues are *independent* of their initial orders in the range $[HB(\alpha_L), OU(\alpha_H)]$: If demand is low, they price to sell $HB(\alpha_L)$ and have leftover inventory; if demand is high, they reorder up to and sell $OU(\alpha_H)$. Therefore, the firms' preferences over initial order quantities in this range only depend on the relative cost of ordering early, at unit cost c, or later at unit cost C. The firms weakly prefer ordering initially at most $HB(\alpha_L)$ if $c \geq C/2$, and at least $OU(\alpha_H)$ if $c \leq C/2$.

Figure 2.3 illustrates, for $\gamma = 0.7$, how the conditions in Proposition 2.3 partition the parameter space of market size ratios r_α and order cost ratios r_c into three regions, each corresponding to one of the three possible second-stage price-reorder strategies. For $\gamma = 0.7$, Part (i) of Proposition 2.3 applies for market size ratios $r_\alpha \leq 1.32$, Part (ii) for $r_\alpha \in (1.32, 2)$, and Part (iii) for $r_\alpha \geq 2$.

Part (i). If the high- and low-demand markets are of sufficiently similar size ($r_\alpha \leq 1.32$), firms never end up with excess inventory. Their initial orders are moderate, such that they prefer to clear their inventories even under low demand. The order cost ratio only affects whether the firms make use of reorder flexibility to delay part of their order; this is the case only if reordering is sufficiently cheap, i.e., the order cost ratio r_c exceeds the threshold $\overline{r}_c(r_\alpha,\gamma)$.

If the high- and low-demand market sizes differ more significantly ($r_\alpha > 1.32$), firms are willing to order more aggressively initially, at the risk of overstocking under low demand – provided that initial ordering is cheap enough, i.e., the order cost ratio r_c is low enough.

Figure 2.3: Proposition 2.3: Symmetric R Game Equilibrium as a Function of Market Size Ratio and Order Cost Ratio ($\gamma = 0.7$)



Part (ii). For moderately different market sizes ($1.32 < r_\alpha < 2$), the equilibrium strategies depend on two thresholds on the order cost ratio. If the order cost ratio r_c is below the lower threshold $\underline{r}_c(r_\alpha, \gamma)$, the firms initially order enough for high demand, but more than they wish to sell under low demand. If the order cost ratio r_c exceeds the larger threshold $\overline{r}_c(r_\alpha, \gamma)$, the firms initially order so little that they reorder under high demand, but they price to clear their inventories under low demand. If the order cost ratio r_c is in the intermediate range $[\underline{r}_c(r_\alpha, \gamma), \overline{r}_c(r_\alpha, \gamma)]$, the firms' initial orders are low enough so they price to clear inventories under low demand, yet large enough so they do not reorder under high demand.

Part (iii). For sufficiently different market sizes ($r_\alpha \geq 2$), firms do not match the demand with their initial order. If early ordering is relatively cheap, that is, $r_c \leq r_c(r_\alpha, \gamma)$, then firms initially order more than they sell under low demand and do not reorder; otherwise, they initially order less than they need under high demand and do reorder if demand is high.

2.5 The Impact of Reorder Flexibility on Orders and Profits

In this section we compare the equilibria with and without reorder flexibility. We call firms inflexible in the N game and flexible in the R game. In Section 2.5.1 we identify two conditions that determine the impact of reorder flexibility on initial orders. In Section 2.5.2 we use these conditions to explain the known results that under *quantity competition*, reorder flexibility reduces initial orders and improves expected profits. These result are

in stark contrast to ours: In Section 2.5.3 we show that under *price competition*, reorder flexibility may increase initial orders and lower profits, and moreover, a reorder option cannot benefit firms if products are sufficiently close substitutes.

2.5.1 Two Conditions Determine the Impact of Reorder Flexibility on Orders

The availability of a reorder option has two effects on the initial equilibrium orders of flexible firms, in comparison to the equilibrium order quantity x^{N*} of inflexible firms. First, reorder flexibility softens the firms' output constraints from their initial procurements, which may intensify their second-stage competition. Second, reorder flexibility allows firms to reduce overstocking risks in matching supply with demand. The first effect may give flexible firms an incentive to sell more than x^{N*} in the second stage *if* demand is high, and the magnitude of the second effect determines in such cases whether they *initially* order more or less than x^{N*}.

To make this discussion precise, consider first the N game equilibrium orders. The inflexible firms hedge their bets between low and high demand, ordering up to the point where their expected marginal second-stage equilibrium revenue equals the initial unit procurement cost:

$$\frac{1}{2}\left(\frac{\partial \pi_i^{N*}\left(\mathbf{x}^{N*};\alpha_L\right)}{\partial x_i}+\frac{\partial \pi_i^{N*}\left(\mathbf{x}^{N*};\alpha_H\right)}{\partial x_i}\right)=c. \tag{2.8}$$

As a result, they order more than optimal for known low demand and less than optimal for known high demand. From (2.8), we have

$$c-\frac{\partial \pi_i^{N*}\left(\mathbf{x}^{N*};\alpha_L\right)}{\partial x_i}=\frac{\partial \pi_i^{N*}\left(\mathbf{x}^{N*};\alpha_H\right)}{\partial x_i}-c>0, \tag{2.9}$$

i.e., the marginal profit loss if demand is low (the LHS) equals the marginal profit gain if it is high.

The flexible firms initially order more than x^{N*} units, if and only if two conditions hold:

(A) The flexible firms want to sell more than x^{N*} units under high demand. That is, x^{N*} is smaller than the high-demand order-up-to level:

$$x^{N*}<OU\left(\alpha_H\right). \tag{2.10}$$

This condition holds if and only if the marginal reorder cost C is lower than a flexible firm's high-demand marginal revenue at x^{N*} from unilaterally dropping its price. This marginal revenue exceeds an inflexible firm's high-demand marginal revenue at x^{N*} from unilaterally increasing its inventory (the term $\partial \pi_i^{N*}\left(\mathbf{x}^{N*}; \alpha_H\right)/\partial x_i$ in (2.8)). In the degenerate case without demand uncertainty, i.e., $\alpha_L = \alpha_H$, whenever condition *(A)* holds, the flexible firms initially order more than x^{N*} units and have lower profits than inflexible firms. However, under stochastic demand, firms must also evaluate under- and overstocking costs.

(B) Expected marginal understocking costs at x^{N*} exceed expected marginal overstocking costs. Ordering initially more than x^{N*} units, at a lower cost, realizes procurement cost savings only if condition *(A)* holds *and* demand turns out to be high; otherwise, if demand turns out to be low, more initial inventory yields lower prices and profits. Therefore, flexible firms initially order more than x^{N*} if, and only if, *(A)* holds and the expected marginal cost saving from early procurement under high demand exceeds the expected marginal profit loss at x^{N*} under low demand:[4]

$$C - c > c - \frac{\partial \pi_i^{N*}\left(\mathbf{x}^{N*}; \alpha_L\right)}{\partial x_i}. \tag{2.11}$$

If condition *(A)* is violated, then the flexible firms initially order the same amount as the inflexible firms, x^{N*}, and do not reorder in the second stage. If condition *(A)* holds but *(B)* is violated, then the flexible firms initially order (weakly) *less than* x^{N*} and their order-up-to level under high demand strictly exceeds x^{N*}. An important implication is that, whenever the equilibrium orders with reorder flexibility differ from those without, the flexible firms end up with *more inventory under high demand* than inflexible firms with a single order. As shown in Section 2.5.3, this over-ordering hurts the flexible firms' profits under high demand. Table 2.1 summarizes this discussion.

Table 2.1: Comparison of Equilibrium Order Strategies in N and R Games.

Flexible Firms Order Initially		Conditions
Same:	$x^{R*} = x^{N*}$	$x^{N*} \geq OU\left(\alpha_H\right)$
More:	$x^{R*} > x^{N*}$	$x^{N*} < OU\left(\alpha_H\right)$ and $C - c > c - \frac{\partial \pi_i^{N*}\left(\mathbf{x}^{N*}; \alpha_L\right)}{\partial x_i}$
Less:	$x^{R*} \leq x^{N*}$	$x^{N*} < OU\left(\alpha_H\right)$ and $C - c \leq c - \frac{\partial \pi_i^{N*}\left(\mathbf{x}^{N*}; \alpha_L\right)}{\partial x_i}$

[4]Since low and high demand are equally likely, we omit the probabilities from this expression.

2.5.2 Quantity Competition: Reorder Flexibility Reduces Initial Orders, Improves Profits

Before we discuss the effects of reorder flexibility in our model with price competition, consider the case where second-stage prices and reorder quantities are determined under *quantity competition*. This mode of competition yields the same results as for the monopoly case ($\gamma = 0$): Flexible firms initially order (weakly) less than inflexible firms. This follows because conditions *(A)* and *(B)* are mutually exclusive. Namely, if it is profitable under high demand to order more than the optimal no-reorder quantity x^{N*}, then it is cheaper to delay doing so until the second stage. Mathematically, under quantity competition (and for a monopoly) condition *(A)* is equivalent to

$$\frac{\partial \pi_i^{N*}\left(\mathbf{x}^{N*}; \alpha_H\right)}{\partial x_i} > C,$$

which, together with (2.9), implies that condition *(B)* cannot hold:

$$c - \frac{\partial \pi_i^{N*}\left(\mathbf{x}^{N*}; \alpha_L\right)}{\partial x_i} = \frac{\partial \pi_i^{N*}\left(\mathbf{x}^{N*}; \alpha_H\right)}{\partial x_i} - c > C - c.$$

That is, the marginal profit loss under low demand exceeds the cost savings from early procurement.

This argument sheds light on the finding in Lin and Parlaktürk (2012) who consider a reorder option for duopoly retailers that sell a homogeneous product under *quantity competition*, in contrast to our R game under price competition. In their model competition between "fast" retailers that can reorder after learning demand yields (weakly) smaller initial orders and larger expected retailer profits, compared to competition between "slow" retailers without a reorder option.

2.5.3 Price Competition: Reorder Flexibility May Increase Initial Orders, Hurt Profits

The results under quantity competition are in stark contrast to ours under price competition. As we discuss in this section, the R game may yield larger initial orders and lower expected profits than the N game, and moreover, a reorder option cannot benefit firms if products are sufficiently close substitutes. The following proposition summarizes these results.

Proposition 2.4 (Effects of Reorder Flexibility). *Assume* $\alpha_L = C(1-\gamma)$.

Consider the expected profits and the order strategies under the Pareto-dominant symmetric equilibrium of the "no-reorder" game and of the "reorder" game.

1. *(Value of reorder flexibility). The firms with reorder flexibility are not more profitable than those without, except if the following three conditions hold:*

   (a) *the products are sufficiently differentiated, i.e.,* $\gamma < 0.875$*;*

   (b) *the market size variability is moderate, i.e.,* $\underline{r}_\alpha(\gamma) < r_\alpha < \overline{r}_\alpha(\gamma)$*, where* $\overline{r}_\alpha(\gamma) < \infty$ *for* $\gamma > 0$*;*

   (c) *reordering is relatively inexpensive, i.e.,* $\underline{r}_c(r_\alpha, \gamma) < r_c \leq 1$*.*

2. *(Order strategies when reorder flexibility is valuable). Whenever the firms with reorder flexibility are more profitable than inflexible firms, they order less initially, but under high demand they reorder and sell more inventory, than the inflexible firms, i.e.,* $x^{R*} < x^{N*} < OU(\alpha_H)$*.*

3. *(Equal equilibrium outcomes). The firms with reorder flexibility order the same amount as those without, and they do not reorder, i.e.,* $x^{R*} = x^{N*} \geq OU(\alpha_H)$*, if and only if:*

   (a) *the market size ratio is below a threshold, i.e.,* $r_\alpha \leq \overline{\overline{r}}_\alpha(\gamma)$*, where* $\overline{\overline{r}}_\alpha(\gamma) < \infty$ *for* $\gamma > 0$*;*

   (b) *reordering is relatively expensive, i.e.,* $r_c \leq \overline{r}_c(r_\alpha, \gamma)$*.*

By part 2. of Proposition 2.4, reorder flexibility always hurts profits if the flexible firms order more ex ante than inflexible firms; the additional inventory increases procurement costs and lowers prices. However, reorder flexibility may also hurt profits in cases where the flexible firms order less initially, compared to inflexible firms; the flexible firms are only better off under the additional conditions 1.(a)-(c) of Proposition 2.4. Next we compare the orders of flexible and inflexible firms. Then we discuss the conditions under which reorder flexibility improves firm profits. In this discussion, we say "equilibrium" as shorthand for "Pareto-dominant symmetric equilibrium".

Impact of reorder flexibility on equilibrium orders. Consider the impact of procurement costs on the flexible firms' initial order incentives, relative to the N game equilibrium. An increase in the order cost ratio r_c has two effects on the conditions *(A)* and *(B)* of Section 2.5.1.

1. It reduces the flexibility cost, which creates a stronger incentive for the flexible firms to sell more than x^{N*} units under high demand. That is, *(A)* is likelier to hold.

2. It reduces the early procurement cost savings under high demand and increases the expected marginal overstocking costs under low demand, making it more attractive for the flexible firms to order less initially and reorder only under high demand. That is, *(B)* is likelier to be violated.

Figure 2.4 shows the impact of reorder flexibility on equilibrium orders for $\gamma = 0.7$. It partitions the parameter space of market size ratios r_α and order cost ratios r_c into three regions, depending on whether the flexible firms initially order the same as ($x^{R*} = x^{N*}$), more than ($x^{R*} > x^{N*}$), or less than ($x^{R*} < x^{N*}$), the inflexible firms. These regions correspond to the cases in Table 2.1.

Figure 2.4: Equilibrium Orders ($\gamma = 0.7$)

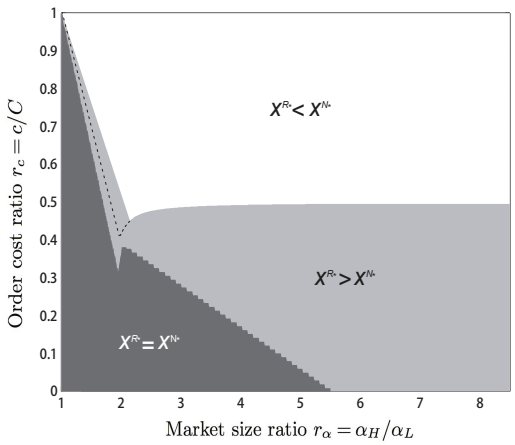
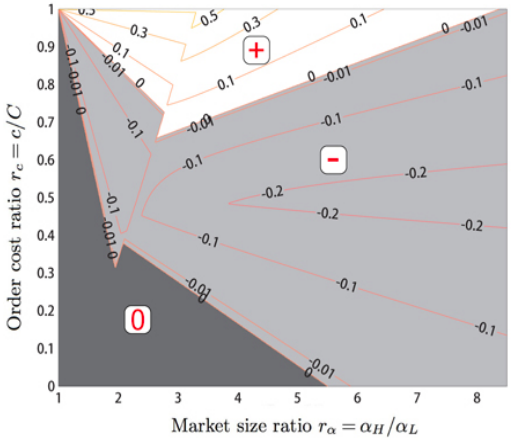


Figure 2.5: Value of Flexibility ($\gamma = 0.7$)

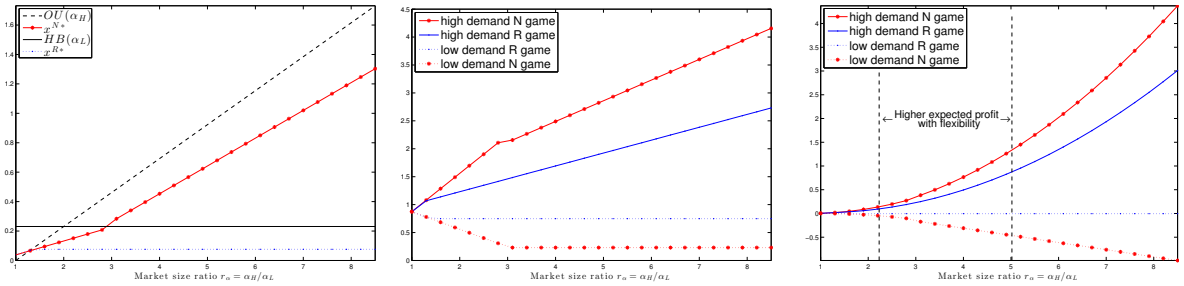


If the high-demand market size is below a threshold ($r_\alpha \leq 5.51$ where $\overline{\overline{r}}_\alpha(\gamma) = 5.51$ for $\gamma = 0.7$ is the threshold in part 3(a) of Proposition 2.4), the flexible firms order the same as, more than, or less than inflexible firms, depending on whether the order cost ratio is low, intermediate, or high, respectively. A low order cost ratio implies a relatively high flexibility cost, so flexible and inflexible firms order the same amounts; see part 3(b) of Proposition 2.4. For an intermediate order cost ratio, the flexibility cost is such that flexible firms want to sell more than x^{N*} under high demand, but delaying this order is too costly, so they order more upfront. For a sufficiently high order cost ratio, the flexibility cost and the early procurement cost savings are negligible, so that flexible firms initially order less than inflexible firms, and more later if needed.

If the high-demand market is sufficiently large ($r_\alpha > 5.51$ for $\gamma = 0.7$), the flexible firms' initial orders differ from those of inflexible firms, regardless of the order cost ratio. In this case the high-demand market fosters such intense price competition that the flexible firms want to sell more than x^{N*} units under high demand, even if initial procurement is free ($c = 0$). If the initial cost is below a threshold, they procure the extra units in the first stage. Otherwise, they order less than x^{N*} initially and reorder up to a higher inventory level under high demand.

Figure 2.6a shows, for order cost ratio $r_c = 0.8$, how the equilibrium orders x^{N*} and x^{R*} depend on the market size ratio r_α ($\gamma = 0.7$ as in Figure 2.4). The flexible firms initially order the same as or more than the inflexible firms for $r_\alpha \leq 1.4$, and strictly less for $r_\alpha > 1.4$. Above this level, only the inflexible firms' order x^{N*} increases in the high-demand market size. The flexible firms' initial order x^{R*} stays constant as it balances the marginal profit loss from overstocking under low demand with the marginal cost saving from early procurement under high demand.

Figure 2.6: Numerical Example ($\gamma = 0.7, r_c = 0.8$)



(a) Quantities (b) Prices (c) Conditional Profits

Value of flexibility: impact of reorder flexibility on equilibrium profits. By part 2. of Proposition 2.4, in cases where firms benefit from reorder flexibility, they order less ex ante and more ex post under high demand, compared to inflexible firms. This condition is only necessary, as noted above. The flexible firms only gain higher expected profits under the additional conditions in part 1 of Proposition 2.4: (a) products are sufficiently differentiated, (b) the market size variability is moderate, and (c) reordering is relatively inexpensive. Figure 2.5 illustrates these conditions for $\gamma = 0.7$. It partitions the parameter space of market size ratios r_α and order cost ratios r_c into three regions, one where reorder flexibility has no profit effect (part 3. of Proposition 2.4), one where it hurts expected profits, and one where it increases expected profits (part 1. of Proposition 2.4). The region of positive reorder benefit is significantly smaller than the set of all (r_α, r_c)-pairs where flexible firms order less ex ante than in the N game (shown in Figure 2.4). Conditions 1.(a)-(c) of Proposition 2.4 result from two countervailing profit effects

of reorder flexibility:

1. *Downside protection under low demand.* The flexible firms are better off than the inflexible firms under low demand. Keeping their initial inventory low allows them to charge higher prices (see Figure 2.6b) and eliminate leftover inventory, compared to the inflexible firms (see Figure 2.6a). The value of this downside protection increases in the market size ratio r_α (see Figure 2.6c) and in the order cost ratio r_c: The more disparate the market sizes and the more expensive initial procurement, the more valuable the reorder option. If the order cost ratio r_c is sufficiently low, the flexible firms' gains from downside protection under low demand are too small to offset any losses under high demand, regardless of other factors.

2. *Intensified competition under high demand.* The flexible firms are worse off than the inflexible firms under high demand. If reordering is relatively inexpensive, the flexible firms over-order to a larger inventory level ($OU(\alpha_H) > x^{N*}$ as shown in Figure 2.6a) and compete more aggressively in price, compared to the inflexible firms (see Figure 2.6b), so they have larger procurement volumes and unit costs, and lower prices. This profit loss increases in the market size ratio r_α (Figure 2.6c) and in the product substitution parameter γ: both effects foster more intense competition. If the products are insufficiently differentiated ($\gamma \geq 0.875$), the flexible firms' losses under high demand exceed any gains under low demand, regardless of other factors.

These two profit effects explain why reorder flexibility benefits firms only if the market size variability is moderate (condition 1.(b) of Proposition 2.4). If the high- and low-demand markets are of similar size, the value of reorder flexibility for downside protection is insignificant because even inflexible firms price to clear their inventory under low demand. The value of downside protection increases in the high-demand market size, as inflexible firms are willing to incur the risk of overstocking under low demand. However, the detrimental effect of over-ordering also increases in the high-demand market size. Therefore, only for intermediate high-demand market size (in Figure 2.6c for $r_\alpha \in (2.2, 5.0)$) does the profit gain from downside protection under low demand exceed the profit loss from intensified competition under high demand.

Summary. Reorder flexibility under *price competition* may lead to smaller or larger initial orders. Flexible firms generate higher profits only if they order less initially, and in addition, products are sufficiently differentiated, the market size variability is moderate, and reordering is relatively inexpensive. Otherwise, reorder flexibility *hurts* profits. As

discussed in Section 2.3.2, under volume flexibility the conditions for the equivalence of price and quantity competition may be violated, and price competition may be a more appropriate model of price formation. Our results are in stark contrast to those under *quantity competition*: In that case reorder flexibility consistently yields lower initial orders and higher profits, as in the monopoly case. These contrasting results underscore the importance of understanding how flexibly firms can increase their supply, in order to better predict and improve performance under reorder flexibility.

2.6 Flexibility Selection: Unilateral Flexibility is Not An Equilibrium

We have so far restricted attention to symmetric flexibility configurations. In this section we justify this focus: We show that in the flexibility-selection stage that precedes the procurement-pricing decisions, unilateral reorder flexibility is not an equilibrium. In Section 2.6.1 we characterize the initial order equilibria under unilateral reorder flexibility. In Section 2.6.2 we characterize the flexibility-selection equilibria and explain why unilateral reorder flexibility is not an equilibrium. In Section 2.6.3 we highlight the profit implications of the symmetric flexibility equilibria.

2.6.1 Initial Order Equilibria Under Unilateral Reorder Flexibility

Consider the two-stage procurement-pricing decisions under *unilateral* reorder flexibility, referred to as the U game. Both firms place initial orders before, but only one firm has the option to reorder after observing the market size; as before both firms have price flexibility. We call the firm who can reorder *flexible* and the one who cannot *inflexible*. Let $\mathbf{x}^{U*} = (x_I^{U*}, x_F^{U*})$ denote equilibrium initial orders in the U game, where x_I^{U*} and x_F^{U*} are the orders of the inflexible and flexible firm, respectively. The following result establishes that there exists at least one initial order equilibrium.

Proposition 2.5 (First Stage Order Equilibria: U Game). *Consider the U game with $\alpha_L = C(1-\gamma) < \alpha_H$. There exists at least one initial order equilibrium $\mathbf{x}^{U*}$. There are two thresholds $\underset{\sim}{r}_c(r\alpha, \gamma) < \tilde{r}_c(r\alpha, \gamma)$ on the order cost ratio such that:*

(i) if $r_c \leq \underset{\sim}{r}_c(r\alpha, \gamma)$, there is a unique equilibrium and it is symmetric. Moreover, $\mathbf{x}^{U*} = \mathbf{x}^{R*} = \mathbf{x}^{N*}$;

(ii) *if* $\underline{r}_c(r\alpha,\gamma) < r_c < \tilde{r}_c(r\alpha,\gamma)$, *there is a continuum of equilibria. Moreover,* $x_F^{U*} > x^{N*}$;

(iii) *if* $\tilde{r}_c(r\alpha,\gamma) \le r_c \le 1$, *there is a unique equilibrium and it is asymmetric. Moreover,* $x_F^{U*} < x^{R*} < x^{N*}$.

Figure 2.7 illustrates Proposition 2.5 for $\gamma = 0.7$ and shows under what conditions the flexible firm in the U game initially orders as much as ($x_F^{U*} = x^{N*}$), more than ($x_F^{U*} > x^{N*}$), or less than ($x_F^{U*} < x^{N*}$) the firms in the N game. As seen by comparison with Figure 2.4, unilateral and bilateral reorder flexibility have similar effects on the flexible firm's initial order.

Figure 2.7: Equilibrium Orders of U Game ($\gamma = 0.7$)

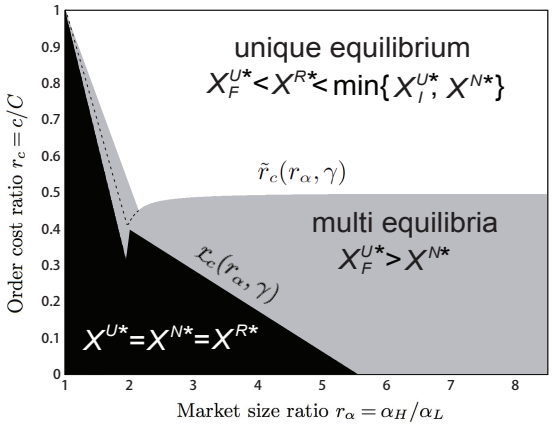


Figure 2.8: Payoff Matrix of Flexibility Strategies

Firm 1 \ Firm 2	N	R
N	Π^{N*}, Π^{N*}	Π_I^{U*}, Π_F^{U*}
R	Π_F^{U*}, Π_I^{U*}	Π^{R*}, Π^{R*}

If the order cost and market size ratios are sufficiently low (part (i) of Prop. 2.5: $r_c \le \underline{r}_c(r_\alpha,\gamma)$ and $r_\alpha \le 5.51$ for $\gamma = 0.7$, the black region in Figure 2.7), the firm with unilateral reorder flexibility has no incentive to sell more than x^{N*} under high demand, so that the U game equilibrium is the same as in the N and R games. Otherwise, the flexible firm has an incentive to sell more than x^{N*} under high demand. If the order cost ratio is in some intermediate range (part (ii) of Prop. 2.5: $\underline{r}_c(r_\alpha,\gamma) < r_c < \tilde{r}_c(r_\alpha,\gamma)$, the grey region in Figure 2.7), deferring the procurement is too costly, so the flexible firm orders all inventory upfront ($x_F^{U*} > x^{N*}$). However, for sufficiently high order cost ratio (part (iii) of Prop. 2.5: $r_c \ge \tilde{r}_c(r_\alpha,\gamma)$, the white region in Figure 2.7), reordering is so cheap that the flexible firm initially orders less than symmetrically inflexible firms ($x_F^{U*} < x^{N*}$) and reorders if demand is high. In this case the flexible firm initially orders even less, whereas the inflexible firm orders more, than under symmetric reorder flexibility, that is, $x_F^{U*} < x^{R*} < x^{N*}$ and $x_I^{U*} > x^{R*}$. This follows because the only way for the inflexible

firm to prepare for potentially high demand is to place a relatively large initial order. In response, the flexible firm reduces its initial order to further mitigate a potential loss under low demand, knowing that it can use the relatively cheap reorder option under high demand.

2.6.2 Reorder Flexibility Selection: Unilateral Flexibility is Not An Equilibrium

Consider the flexibility-selection stage that precedes the procurement-pricing decisions discussed so far. Each firm chooses whether to have reorder flexibility (R) or not (N). Let (R, R) and (N, N) denote the symmetric flexibility strategies, (N, R) and (R, N) the asymmetric strategies, where the first letter refers to firm 1. Our analysis assumes a zero fixed cost for reorder flexibility. This assumption isolates the strategic effects of flexibility selection without biasing the comparison against reorder flexibility by imposing an upfront investment in addition to the higher unit cost. In Section 2.6.3 we explain why our main results are robust if this assumption is relaxed.

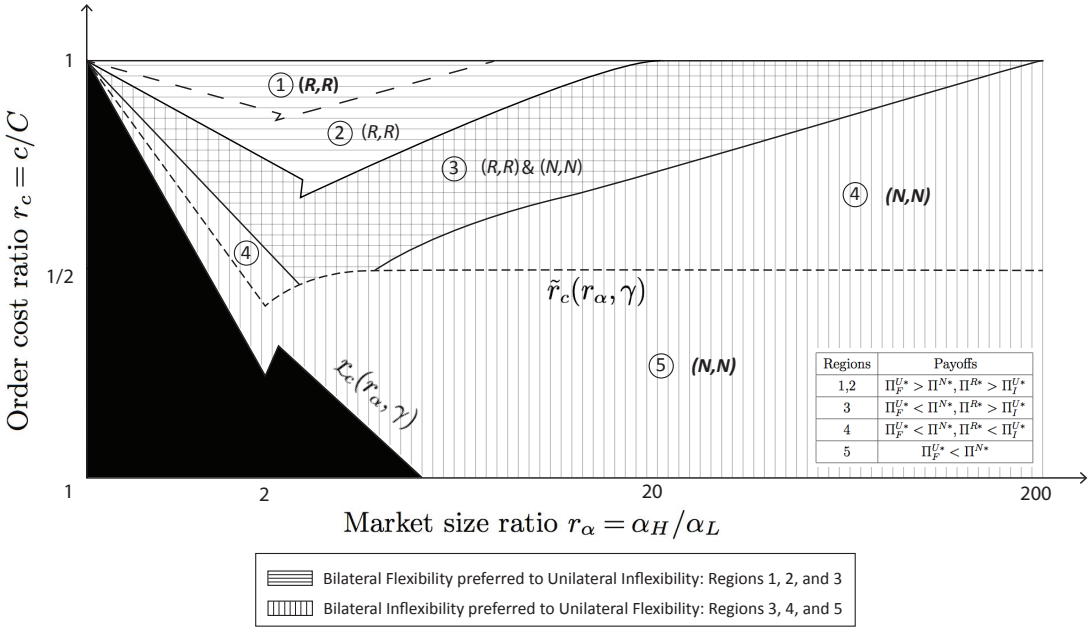
We show that unilateral reorder flexibility is *not* an equilibrium. From the payoff matrix in Figure 2.8, it is easy to see that the asymmetric flexibility strategies (R, N) or (N, R) can be an equilibrium if and only if (1) unilateral flexibility is at least as profitable as bilateral inflexibility (i.e., $\Pi_F^{U*} \geq \Pi^{N*}$), and (2) unilateral inflexibility is at least as profitable as bilateral flexibility (i.e., $\Pi_I^{U*} \geq \Pi^{R*}$). We find that these conditions cannot hold. By part (i) of Proposition 2.5, all flexibility configurations yield the same equilibrium if $r_c \leq \underset{\sim}{r}_c(r_\alpha, \gamma)$. Proposition 2.6 establishes for a large set of cases (part (iii) of Prop. 2.5: $\tilde{r}_c(r_\alpha, \gamma) \leq r_c \leq 1$) that conditions (1) and (2) are mutually exclusive. Numerical results confirm for the remaining cases (part (ii) of Prop. 2.5: $\underset{\sim}{r}_c(r_\alpha, \gamma) < r_c < \tilde{r}_c(r_\alpha, \gamma)$) that condition (1) is violated, that is, bilateral inflexibility is strictly more profitable than unilateral flexibility (i.e., $\Pi^{N*} > \Pi_F^{U*}$).

Proposition 2.6 *If $\tilde{r}_c(r\alpha, \gamma) \leq r_c \leq 1$, then asymmetric flexibility strategies cannot be an equilibrium in the reorder flexibility selection game.*

Figure 2.9 illustrates the results for $\gamma = 0.7$ (Proposition 2.6 applies to the regions 1-4, the numerical analysis to region 5).

(i) Unilateral reorder flexibility is strictly more profitable than bilateral inflexibility (i.e., $\Pi_F^{U*} > \Pi^{N*}$), if and only if the order cost ratio r_c is sufficiently high and the market size ratio r_α is in an intermediate range. However, in this case the firms also prefer bilateral flexibility over unilateral inflexibility (i.e., $\Pi^{R*} > \Pi_I^{U*}$). That is, selecting reorder

Figure 2.9: Preferences over Flexibility Configurations ($\gamma = 0.7$; Figure is Re-Scaled).



flexibility is the dominant strategy. The parameter conditions for this case correspond to regions 1 and 2 in Figure 2.9. These preference conditions for unilateral reorder flexibility over bilateral inflexibility parallel those for bilateral reorder flexibility to increase profits over bilateral inflexibility (part 1 of Proposition 2.4), and the underlying intuition discussed in Section 2.5.3 applies here as well: First, reorder flexibility can be of value only if reordering is so cheap (i.e., the order cost ratio is so high) that the flexible firm reduces its initial order and reorders only under high demand. Second, only for an intermediate high-demand market size does reorder flexibility yield a sufficiently high profit gain from downside protection under low demand to offset the loss from intensified competition under high demand.

(*ii*) Otherwise, the firms strictly prefer bilateral inflexibility to unilateral reorder flexibility (i.e., $\Pi^{N*} > \Pi_F^{U*}$), see regions 3-5 in Figure 2.9. The intuition for these cases parallels that for bilateral reorder flexibility to reduce profits relative to bilateral inflexibility (refer to the discussion of Figures 2.4-2.5 in Section 2.5.3). In regions 3 and 4, the firm with unilateral flexibility initially orders less than in the N game (i.e., $x_F^{U*} < x^{N*}$ by part (*iii*) of Prop. 2.5) and therefore enjoys profit gains from downside protection under low demand, but these are overwhelmed by losses from intensified competition under high demand. In region 5, however, the firm with unilateral flexibility enjoys no downside protection, because it initially orders *more* than in the N game (i.e., $x_F^{U*} > x^{N*}$

by part (*ii*) of Prop. 2.5). As a result, it faces higher procurement costs and lower prices than in the N game.

2.6.3 Symmetric Flexibility Equilibria: Profit Implications

Profits are as follows under the symmetric flexibility equilibria (refer to Figure 2.9).

In regions 1 and 2, selecting reorder flexibility is the *dominant* strategy, so that bilateral reorder flexibility (R, R) is the unique equilibrium. However, by Proposition 2.4, the resulting expected profits Pareto-dominate those under bilateral inflexibility (i.e., $\Pi^{R*} > \Pi^{N*}$) only in region 1, whereas the *opposite* holds in region 2 (and also in regions 3-5). Therefore, parameters in region 2 give rise to the *worst-case scenario*: Bilateral reorder flexibility is the unique equilibrium but both firms are better off under bilateral inflexibility (i.e., $\Pi^{N*} > \Pi^{R*}$). In this case, which parallels the traditional "prisoner's dilemma", reordering is so cheap that firms cannot credibly commit to inflexibility under price competition, yet still so expensive that the losses under high demand exceed the gains under low demand. As a result, each firm individually prefers reorder flexibility (i.e., $\Pi_F^{U*} > \Pi^{N*}$ and $\Pi^{R*} > \Pi_I^{U*}$) although the outcome hurts their profits.

In regions 3-5, bilateral inflexibility is the Pareto-dominant equilibrium. In region 3, bilateral flexibility is also an equilibrium, but profits are higher without reorder option (Proposition 2.4).

To summarize, reorder flexibility benefits firms only under fairly restrictive conditions (region 1). Absent these conditions, firms can commit to inflexibility and avoid the downside of reorder flexibility only in some cases (regions 3-5); in other cases price competition compels them to have reorder flexibility even though it hurts their profits (region 2).

Effect of fixed cost for reorder flexibility. Our analysis assumes a zero fixed cost for reorder flexibility. The presence of a positive fixed cost (or a flat-fee as part of the procurement tariff) would not alter our key findings about the downside of reorder flexibility: The main effect of a fixed cost is that reorder flexibility becomes *less*, not more, attractive. Therefore, the set of parameters for which firms prefer unilateral reorder flexibility to bilateral inflexibility (i.e., $\Pi_F^{U*} > \Pi^{N*}$) is smaller if there is a fixed cost. That is, regions 1 and 2 shrink, whereas regions 3-5 expand. An asymmetric flexibility equilibrium may potentially arise in a subset of these reduced regions 1 and 2. However, the worst-case scenario, whereby choosing reorder flexibility is the dominant strategy even though firms are more profitable with bilateral inflexibility, will still arise for certain parameters. Furthermore, in the expanded regions 3-5, bilateral inflexibility

would Pareto-dominate bilateral flexibility (i.e., $\Pi^{N*} > \Pi^{R*}$) and outperform unilateral flexibility (i.e., $\Pi^{N*} > \Pi_F^{U*}$) even more strongly.

2.7 Discussion and Concluding Remarks

2.7.1 Discussion

Demand function. Our analysis assumes linear demand for tractability but our main results do not rely on this assumption: With nonlinear demand, reorder flexibility may also yield larger orders under price competition, which is the key driver of profit losses. Furthermore, as noted in Section 2.2 the demand system (2.1) does not explicitly model perfectly substitutable products. However, (2.1) is a rescaling of the alternative form $d_i(\mathbf{p};\alpha) = [(1-\gamma)\alpha - p_i + \gamma p_{-i}]/(1-\gamma^2)$ which does capture perfectly substitutable products and also yields the same managerial insights. First, all comparisons of reorder configurations for a *fixed* γ in our model remain intact when (2.1) is scaled back to the alternative form. Second, comparative statics on γ obtained in our model remain qualitatively the same as under the alternative system, because there is a one-to-one increasing correspondence between γ in our model and the cross-elasticity $\frac{\gamma}{1-\gamma^2}$ in the alternative demand system.

Binary market size distribution. Relaxing the assumption that the market size has a binary distribution does not change our main results that under price competition, reorder flexibility may increase initial orders and hurt profits. (Lin and Parlaktürk 2012 also assume a two-point market size distribution in their study of reorder flexibility under quantity competition.) Let the market size α follow a general distribution with probability density function $f(\cdot)$ over $[\underline{\alpha}, \overline{\alpha}]$. Following Section 2.5.1, it is easy to see from (2.8) that the N game equilibrium order should satisfy: $\int_{\underline{\alpha}}^{\overline{\alpha}} \frac{\partial \pi_i^{N*}(\mathbf{x}^{N*};\alpha)}{\partial x_i} f(\alpha) d\alpha = c$. For any market size α such that $OU(\alpha) > x^{N*}$, the flexible firms want to sell more than x^{N*}. They will over-order ex ante or ex post for these market size realizations, depending on the costs and outcome probabilities (generalizing the conditions in Table 2.1), which intensifies price competition and hurts their profits. For a sufficiently dispersed market size distribution $f(\cdot)$, the incidence of such larger market sizes and the resulting losses would offset the gains from downside protection under lower demand.

Information about initial orders. The assumption that firms know their competitor's past supply decisions (i.e., initial orders in our model) is standard, both in the flexibility literature and in dynamic models of inventory competition (e.g., Van Mieghem and Dada 1999, Netessine et al. 2006, Anupindi and Jiang 2008, Olsen and Parker 2008, Caro and

Martínez-de-Albéniz 2010, Lin and Parlaktürk 2012). In practice it is not uncommon for third-party watchdogs who scout industry news to disseminate information about firms' orders and inventory positions.

2.7.2 Conclusion

This paper provides the first analysis of reorder flexibility under price competition and identifies its downside: higher initial orders and lower profits. These results are contrary to prior findings under fixed prices or quantity competition. Unilateral reorder flexibility is not an equilibrium. Furthermore, firms can commit to bilateral inflexibility and avoid the downside of reorder flexibility only in some of the cases where it hurts profits. In other cases firms are trapped in a prisoner's dilemma, whereby reorder flexibility is the dominant strategy even though it hurts their profits.

Our results have several implications for marketing and operations. To reap the benefits and avoid the downside of reorder flexibility, firms need to better understand its effects on competition and profitability. First, firms must differentiate their products sufficiently from their competitors. In this sense, reorder flexibility and product innovation must be viewed and managed as complementary capabilities. Second, the detrimental effect of reorder flexibility through intensified price competition depends on factors that determine how flexibly firms can increase their supply. Limitations on upside volume flexibility, such as convex reordering costs or hard capacity constraints, can help mitigate its detrimental effect, but they also reduce the ability of firms to capitalize on demand surges. It is therefore of strategic importance for firms to determine levels of reactive capacity which appropriately balance competitive considerations with the upside and downside risks due to stochastic demand. Third, although the specific pricing and ordering prescriptions from our pre-season replenishment model do not directly transfer to settings with in-season replenishment, our main findings should continue to hold, because in such settings reorder flexibility under price competition may also yield larger orders.

We conclude by outlining some important research avenues. (1) In our analysis firms base their second-stage decisions on a perfect demand signal. If firms receive a noisy demand signal, they need to balance over- and under-stocking risks in their second-stage decisions. (2) It would be similarly interesting to relax the standard assumption of perfect initial-order information and consider how noisy information on competitor inventories affects the results. (3) By endogenizing product differentiation decisions one could study the interplay of strategic product positioning, pricing, and reorder flexibility selection.

(4) We assume that firms are symmetric in terms of their demand and costs. Accounting for asymmetric firms is a challenging opportunity for future research. (5) Studying the sensitivity of our results to different demand uncertainty models may yield additional insights (cf. Anupindi and Jiang 2008). (6) We model substitutable products. It would be interesting to compare our results with those for complementary products.

2.8 Appendix: Proofs

Proof of Lemma 2.1. (R) We start with the R game. First, we characterize firm i's best response given its competitor's price p_{-i} and its own initial inventory x_i. Second, we use this result to solve for the equilibrium prices for each strategy pair. Lastly, we use the equilibrium prices to identify the valid (x_i, x_{-i}) regions, i.e., the partition in Figure 2.1b, corresponding to the strategy pairs. The results of the N game can be considered as a special case of the reorder game where $C = \infty$.

First, from problem (2.4), we see that the marginal profit with respect to p_i is

$$\frac{\partial \pi_i^R}{\partial p_i} = d_i(\mathbf{p};\alpha) + p_i \frac{\partial d_i}{\partial p_i} - C \cdot \mathbf{1}_{\{d_i(\mathbf{p};\alpha) > x_i\}} \frac{\partial d_i}{\partial p_i} = \alpha - 2p_i + \gamma p_{-i} + C \cdot \mathbf{1}_{\{p_i < p_i^c(p_{-i}, x_i)\}},$$

where $p_i^c(p_{-i}, x_i) = \alpha + \gamma p_{-i} - x_i$ is the clearance price, and hence $\pi_i^R(\mathbf{p}, x_i; \alpha)$ is concave in p_i. Equating the marginal profit to zero yields the hold back price $p_i^h(p_{-i}) = (\alpha + \gamma p_{-i})/2$ or the reorder price $p_i^r(p_{-i}) = (\alpha + \gamma p_{-i} + C)/2$.

Second, we solve for all possible equilibrium prices. Note that the best response price $p_i^{R*}(x_i, p_{-i})$ can be expressed in a general form as $p_i^{R*}(x_i, p_{-i}) = m_i(k_i + \gamma p_{-i})$. The choices of m_i and k_i depend on which of the three potential best-response prices is the optimal one. For example, $m_i = 1$ and $k_i = \alpha - x_i$ if $p_i^{R*}(x_i, p_{-i}) = p_i^c(x_i, p_{-i})$. Thus, the equilibrium prices can be solved from the following system of linear equations:

$$\begin{cases} p_i^{R*}(x_i) &= m_i(k_i + \gamma p_{-i}^{R*}(x_{-i})) \\ p_{-i}^{R*}(x_{-i}) &= m_{-i}(k_{-i} + \gamma p_i^*(x_i)) \end{cases} \iff \begin{cases} p_i^{R*}(x_i) &= m_i(k_i + m_{-i}k_{-i}\gamma)/(1 - m_i m_{-i}\gamma^2) \\ p_{-i}^{R*}(x_{-i}) &= m_{-i}(k_{-i} + m_i k_i \gamma)/(1 - m_i m_{-i}\gamma^2) \end{cases}. \tag{2.12}$$

Table 2.2 lists the choices of m_i, m_{-i}, k_i, and k_{-i} for each strategy pair. Substituting the appropriate coefficients back into system (2.12), we can obtain all possible equilibrium prices.

Lastly, we identify region boundaries in the initial inventory space for each specific strategy pair to arise as an equilibrium. Note that a specific price pair is indeed an equilibrium if and only if the initial inventory levels are in a certain region as illustrated in Figure 2(b). For any fixed competitor's initial inventory level x_{-i}, a horizontal line with intercept

Table 2.2: Coefficients of Best-Response Prices

	$R(r,r)$	$R(r,c)$	$R(r,h)$	$R(c,r)$	$R(c,c)$	$R(c,h)$	$R(h,r)$	$R(h,c)$	$R(h,h)$
m_i	$1/2$	$1/2$	$1/2$	1	1	1	$1/2$	$1/2$	$1/2$
m_{-i}	$1/2$	1	$1/2$	$1/2$	1	$1/2$	$1/2$	1	$1/2$
k_i	$\alpha+C$	$\alpha+C$	$\alpha+C$	$\alpha-x_i$	$\alpha-x_i$	$\alpha-x_i$	α	α	α
k_{-i}	$\alpha+C$	$\alpha-x_{-i}$	α	$\alpha+C$	$\alpha-x_{-i}$	α	$\alpha+C$	$\alpha-x_{-i}$	α

x_{-i} intersects with the region boundaries at two points, which serve as the desired thresholds claimed in the stipulation. Using the prices from system (2.12), we can solve for the valid region boundaries. To illustrate, we work with equilibrium strategy pair (r,r) as an example. For this particular strategy pair to be an equilibrium, we require $x_i < (\alpha + \gamma p_{-i}^{R*}(x_{-i}) - C)/2$ and $x_{-i} < (\alpha+\gamma p_i^{R*}(x_i)-C)/2$. Since $p_i^{R*}(x_i)$ and $p_{-i}^{R*}(x_{-i})$ are available from Table 2.2, we have that (r,r) arises as an equilibrium strategy pair if and only if $x_i < OU = (\alpha - C + \gamma C)/(2-\gamma)$. In a similar way, we can determine the boundaries of $R(h,r)$, $R(h,h)$, and $R(r,h)$ as in Figure 2(b). This gives us the other three vertices of $R(c,c)$ as in Figure 2(b): starting from the bottom-right and going counterclockwise, their coordinates are $\left(\frac{\alpha(2+\gamma)+\gamma C}{4-\gamma^2}, \frac{\alpha(2+\gamma)+(\gamma^2-2)C}{4-\gamma^2}\right)$, $\left(\frac{\alpha}{2-\gamma}, \frac{\alpha}{2-\gamma}\right)$, and $\left(\frac{\alpha(2+\gamma)+(\gamma^2-2)C}{4-\gamma^2}, \frac{\alpha(2+\gamma)+\gamma C}{4-\gamma^2}\right)$. One can also easily verify that the lines connecting every pair of adjacent vertices of the diamond $R(c,c)$ give the other boundaries.

(N) In the N game, the potential best-response prices are left with either the hold back price $p_i^h(p_{-i}) = (\alpha + \gamma p_{-i})/2$ or the clearance price $p_i^c(p_{-i}, x_i)$. The equilibrium region boundaries can be similarly determined as in the R game. Equivalently, one can view the region partition generated in the N game as setting C very large in the R game (i.e., the reorder option is too expensive) when the point (OU, OU) goes below the origin and into the third quadrant. ■

Proof of Proposition 2.2. The general idea of the equilibrium proof is straightforward: we identify the best-response functions, and the equilibrium is where the best response functions intersect. To execute this idea, we partition the initial inventory space $\{\mathbf{x} \geq 0\}$ into nine regions dependent on the second-stage equilibrium strategies. We use the second-stage equilibrium strategies to label the regions. For instance, the region $N(\overbrace{\underbrace{c}_{L}\ \underbrace{c}_{H}}^{i}, \overbrace{\underbrace{c}_{L}\ \underbrace{c}_{H}}^{-i})$ consists of all initial inventory vectors $\mathbf{x}$ for which the unique price equilibrium in both the low and high demand scenarios is for both firms to charge the clearance price. We follow the same notation convention for the other regions: the first and second components refer to the second-stage equilibrium strategies for firm

i and $-i$, respectively, and the first and second letters of each component represent the firm's equilibrium strategy under low and high demand, respectively. It is easy to see that, because procurement is costly, it cannot be an equilibrium for firms to hold back inventory under high demand. Therefore a symmetric equilibrium can only be in region $N(cc, cc)$ or $N(hc, hc)$. We solve for an equilibrium candidate by concatenating the two first order conditions (FOCs) of each firm's profit maximization problem for the profit functions corresponding to $N(cc, cc)$ or $N(hc, hc)$. Then we identify conditions under which the candidate is indeed in the intended region and is indeed an equilibrium.

First, we consider region $N(cc, cc)$. The expected profit function of firm i with $\mathbf{x} \in N(cc, cc)$ is $\Pi_i^N(\mathbf{x}) = \left(\frac{\overline{\alpha}}{1-\gamma} - \frac{1}{1-\gamma^2}x_i - \frac{\gamma}{1-\gamma^2}x_{-i}\right)x_i - cx_i$, where $\overline{\alpha} = (\alpha_L + \alpha_H)/2$. For any fixed $x_{-i} \in \left[0, \frac{\alpha_L}{2-\gamma}\right]$, we solve for a line from the FOC of firm i's profit maximization problem: $x_i = \frac{1}{2}\left((1+\gamma)(\overline{\alpha} - c(1-\gamma)) - \gamma x_{-i}\right)$. Since the game is symmetric, these two lines, for $i = 1, 2$, intersect on the diagonal, which yields the symmetric equilibrium candidate $x_l^{N*} = \frac{(1+\gamma)(\overline{\alpha}-c(1-\gamma))}{\gamma+2} \geq 0$, where the inequality is due to $\overline{\alpha} - c(1-\gamma) \geq 0$. To ensure that the equilibrium candidate is indeed in the region $N(cc, cc)$, we require $\hat{\alpha}_H \leq \frac{\gamma^2+\gamma+2}{(1+\gamma)(2-\gamma)}\hat{\alpha}_L + 2 := B_1\hat{\alpha}_L + 2$, where $\hat{\alpha}_L = \frac{\alpha_L}{c(1-\gamma)}$ and $\hat{\alpha}_H = \frac{\alpha_H}{c(1-\gamma)}$. Next, we need to characterize the conditions under which $x_i = x_l^{N*}$ is a best-response in maximizing $\Pi_i^N(\mathbf{x})$ among $x_i \geq 0$ while fixing $x_{-i} = x_l^{N*}$. It is clear that $x_i = x_l^{N*}$ is the expected profit maximizer for any $(x_i, x_{-i} = x_l^{N*}) \in N(cc, cc)$. Moreover, $\Pi_i^N(\mathbf{x})$ with $\mathbf{x} \in N(hh, cc)$ is decreasing in x_i for any fixed x_{-i}. Thus, only the local maximizer of $\Pi_i^N(\mathbf{x})$ with $\mathbf{x} = (x_i, x_{-i} = x_l^{N*}) \in N(hc, cc)$, which is quadratic in x_i, could be a possible best response. It can be shown that if this local maximizer in $N(hc, cc)$ is on the boundary, i.e., $\hat{\alpha}_H < \frac{8+4\gamma-\gamma^3}{8+4\gamma-4\gamma^3-\gamma^3}\hat{\alpha}_L + 2 := m^{**}(\gamma)\hat{\alpha}_L + 2$, it must reside on the boundary between $N(hc, cc)$ and $N(cc, cc)$. By the continuity of the profit at the boundary, such $\mathbf{x}$ with $x_i \in N(hc, cc) \cap N(cc, cc)$ and $x_{-i} = x_l^{N*}$ is dominated by $x_i = x_{-i} = x_l^{N*}$ that is the local maximizer over $N(cc, cc)$. If the local maximizer of $\Pi_i^N(\mathbf{x})$ with $\mathbf{x} = (x_i, x_{-i} = x_l^{N*}) \in N(hc, cc)$ is an interior point of $N(hc, cc)$, i.e., $\hat{\alpha}_H \geq= m^{**}(\gamma)\hat{\alpha}_L + 2$, then $x_i = x_{-i} = x_l^{N*}$ yields equal or more profit than the local maximizer in $N(hc, cc)$ if and only if $K_1\hat{\alpha}_L + 2 \leq \hat{\alpha}_H \leq K_2\hat{\alpha}_L + 2$, where $K_1 := \frac{(\sqrt{2}+1)\gamma^3+(2\sqrt{2}+4)\gamma^2-2\sqrt{2}\gamma-4\sqrt{2}}{(\sqrt{2}+1)\gamma^3+2\sqrt{2}\gamma^2-2\sqrt{2}\gamma-4\sqrt{2}}$ and $K_2 := \frac{(\sqrt{2}-1)\gamma^3+(2\sqrt{2}-4)\gamma^2-2\sqrt{2}\gamma-4\sqrt{2}}{(\sqrt{2}-1)\gamma^3+2\sqrt{2}\gamma^2-2\sqrt{2}\gamma-4\sqrt{2}}$. We verify that $K_1 \leq m^{**}(\gamma) \leq K_2 \leq B_1$ for any given $\gamma \in [0, 1]$, so that $x_i = x_l^{N*}$ is indeed a best response in profit maximization with $x_{-i} = x_l^{N*}$ if and only if $\hat{\alpha}_H \leq K_2\hat{\alpha}_L + 2$. Thus $x_i = x_{-i} = x_l^{N*}$ is a symmetric equilibrium if and only if $\hat{\alpha}_H \leq K_2\hat{\alpha}_L + 2$.

Second, we consider region $N(hc, hc)$. For any fixed $x_{-i} \geq \frac{\alpha_L}{2-\gamma}$, firm i solves the FOC of the expected profit function in $N(hc, hc)$, which yields the line $x_i = \frac{1}{2}\left(\alpha_H(1+\gamma) - \gamma x_{-i}\right) - c(1-\gamma^2)$. This line intersects on the diagonal suggesting the symmetric equilibrium can-

didate $x_i = x_{-i} = x_h^{N*} = \alpha_H \frac{1+\gamma}{2+\gamma} - 2c\frac{1-\gamma^2}{2+\gamma}$, which is equal to or larger than $HB(\alpha_L)$ if and only if $\hat{\alpha}_H \geq B_2\hat{\alpha}_L + 2$, where $B_2 := \frac{2+\gamma}{(2-\gamma)(1+\gamma)}$. Next, we identify the conditions under which $x_i = x_h^{N*}$ is a best response of firm i's profit maximization with $x_{-i} = x_h^{N*}$. Similar to the case in $N(cc, cc)$, there may exist two local maximizers in $N(cc, cc)$ and $N(cc, hc)$ respectively. Following similar procedures, we prove that: (i) $x_i = x_h^{N*}$ yields equal or more profit than the local maximizer in $N(cc, cc)$ if and only if $\hat{\alpha}_H \geq T_1\hat{\alpha}_L + 2$, where $T_1 := \frac{2+\gamma}{4(1+\gamma)}\left(2 + \gamma + \frac{\gamma^2}{2-\gamma}\sqrt{\frac{4-3\gamma^2}{2-\gamma^2}}\right)$, and (ii) $x_i = x_h^{N*}$ yields equal or more profit than the local maximizer in $N(hc, cc)$ if and only if $\hat{\alpha}_H \geq T_2\hat{\alpha}_L + 2$, where $T_2 := \frac{(2+\gamma)(1+\gamma)(2-\gamma)+4\gamma(2+\gamma)\sqrt{1+\gamma}}{(1+\gamma)(-\gamma^2+4\gamma+4)(2-\gamma)}$. Moreover, we verify that $B_2 \leq \max(T_1, T_2)$ for any given $\gamma \in [0, 1]$ (the order of T_1 and T_2 depends on γ). Hence, it follows that $x_i = x_h^{N*}$ is a best response of $\Pi_i^N(\mathbf{x})$ among $x_i \geq 0$ with $x_{-i} = x_h^{N*}$ if and only if $\hat{\alpha}_H \geq \max(T_1, T_2)\hat{\alpha}_L + 2$. Thus, $x_i = x_{-i} = x_h^{N*}$ is a symmetric equilibrium if and only if $\hat{\alpha}_H \geq \max(T_1, T_2)\hat{\alpha}_L + 2$.

Let $m^{**}(\gamma) = K_2$ and $m^*(\gamma) = \max(T_1, T_2)$. We verify that $m^{**}(\gamma) > m^*(\gamma) > 0$ for any given $\gamma \in (0, 1)$. Then we have the desired results on the symmetric equilibria.

Lastly, we eliminate the existence of asymmetric equilibria. We first consider the case that $x_i = x_l^{N*}$ is not a best response to $x_{-i} = x_l^{N*}$. Then, the local maximizer of $\Pi_i^N(\mathbf{x})$ in $N(hc, cc)$ must be the best response, which requires $\hat{\alpha}_H > K_2\hat{\alpha}_L + 2$. To have an asymmetric equilibrium in region $N(hc, cc)$, it is also required that the best response of firm $-i$ is in $N(hc, cc)$ for $\hat{\alpha}_H > K_2\hat{\alpha}_L + 2$. By symmetry, this is equivalent to requiring that the best response of firm i is in $N(cc, hc)$ if $\hat{\alpha}_H > K_2\hat{\alpha}_L + 2$, which contradicts the fact that $x_i = x_{-i} = x_h^{N*} \in N(hc, hc)$ is a symmetric equilibrium if $\hat{\alpha}_H > K_2\hat{\alpha}_L + 2$. Similar arguments apply when $\hat{\alpha}_H \leq K_2\hat{\alpha}_L + 2$ and we can eliminate all the possibilities of asymmetric equilibria. ■

Proof of Proposition 2.3. We call a point on the diagonal of initial order quantities a *symmetric* equilibrium candidate if it is a local maximizer of problem $\max_{x_i \geq 0} \Pi_i^R(\mathbf{x})$. We solve for all symmetric equilibrium candidates and then identify for each candidate conditions under which it is indeed an equilibrium.

Table 2.3: Notation in Proof of Proposition 2.3

$w_0 = (\gamma+2)^2(\gamma^2-2)^2,\ w_0' = (\gamma-2)^{-1}(\gamma+2)^{-1}/2$	$r_{c_2} = w_3 r_\alpha - w_1(\sqrt{w_2 r_\alpha^2 + w_4 r_\alpha + w_5} - w_6)$
$w_1 = (4\gamma^4 - 10\gamma^2 + 8)^{-1} w_0',\ w_2 = (\gamma^2-2)^2 w_0$	$r_{c_5} = -(\rho_2\gamma(1,1,2)r_\alpha + \rho_2\gamma(1,-3,-6))w_0'$
$w_5 = -\rho_6\gamma(7,-24,-32,80,80,-64,-64)w_0$	$r_{c_6} = -(\rho_1\gamma(1,2)r_\alpha + \rho_2\gamma(2,-3,-6))w_0'$
$w_3 = w_0 w_1, w_4 = \rho_6\gamma(2,-20,-2^4,72,2^6,-2^6,-2^6)w_0$	$r_{c_7} = (2\rho_4\gamma(1,-1,-4,2,4))^{-1}\left((-2\gamma^2+\gamma^4)r_\alpha + \right.$
$w_6 = \rho_6\gamma(5,-14,-26,46,60,-40,-48)$	$\rho_3\gamma(-1,-2,2,4) - 2\sqrt{\rho_2\gamma(1,-1,-2)(2-r_\alpha)}$
$r_{c_1} = (-\gamma^2 r_\alpha + 2 + \gamma)w_0',\ r_{c_4} = (r_\alpha - m^{**}(\gamma))/2$	$\left. \cdot\sqrt{\rho_3\gamma(1,1,-1,-2)r_\alpha + \rho_4\gamma(1,-2,-4,2,4)}\ \right)$

First, we consider $r_\alpha \geq 2$, i.e., $\alpha_H \geq 2\alpha_L$. Since $\alpha_L = C(1-\gamma)$, then $0 = OU(\alpha_L) <$

$HB(\alpha_L) \le OU(\alpha_H) < HB(\alpha_H)$. Depending on the relative position of a symmetric point with respect to the HB and OU points of the low and high demand scenarios which determine the second-stage equilibrium strategies, by Table 2.2, the marginal value of initial inventory is

$$\left.\frac{\partial \Pi_i^R(\mathbf{x})}{\partial x_i}\right|_{x_i=x_{-i}} = \begin{cases} \frac{1}{2}\left(\frac{\alpha_L}{1-\gamma} - x_i\frac{2+\gamma}{1-\gamma^2}\right) + \frac{C}{2} - c & \text{if } x_i = x_{-i} \in [0, HB(\alpha_L)], \\ \frac{C}{2} - c & \text{if } x_i = x_{-i} \in (HB(\alpha_L), OU(\alpha_H)), \\ \frac{1}{2}\left(\frac{\alpha_H}{1-\gamma} - x_i\frac{2+\gamma}{1-\gamma^2}\right) - c & \text{if } x_i = x_{-i} \in (OU(\alpha_H), HB(\alpha_H)]. \end{cases}$$

Hence we have the following symmetric equilibrium candidates.

1. For $x_i = x_{-i} \in [0, HB(\alpha_L)]$, setting the derivative to zero yields $x_l^{R*} := \frac{2\left(1-\gamma^2\right)}{2+\gamma} C\left(1 - r_c\right)$. Note that $x_l^{R*} \in [0, HB(\alpha_L)]$ is equivalent to $r_c \in \left[r'_{c_2}, 1\right]$, where $r'_{c_2} := \frac{2\gamma^2-\gamma-2}{2(1+\gamma)(\gamma-2)}$. As in the proof of Proposition 2.2, we need to further identify when x_l^{R*} is indeed a best response while fixing her rival's inventory at $\frac{2\left(1-\gamma^2\right)}{2+\gamma}\left(C - r_c\right)$. After comparing with all possible local optima and noting that $r'_{c_2} \le r_{c_2}$, we conclude that x_l^{R*} is an equilibrium if and only if $r_{c_2} \le r_c \le 1$ when $r_\alpha \ge 2$. Under this equilibrium, firms clear inventory in low demand and reorder in high demand.

2. For $x_i = x_{-i} \in (OU(\alpha_H), HB(\alpha_H)]$, setting the derivative to zero yields $x_h^{R*} := \frac{1-\gamma^2}{2+\gamma} C(r_\alpha - 2r_c)$. Note that $x_h^{R*} > OU(\alpha_H)$ is equivalent to $r_c < r_{c_1}$. Recall that the no-reorder equilibrium is always smaller than $HB(\alpha_H)$ and the profit functions are the same for points along the diagonal between $OU(\alpha_H)$ and $HB(\alpha_H)$ for both R and N games; thus it is implied that $x_h^{R*} < HB(\alpha_H)$. Again, we need to characterize when x_h^{R*} is indeed a best response. We can show that x_h^{R*} is an equilibrium if and only if $r_c < r_{c_1}$ when $r_\alpha \ge 2$. Under this equilibrium, firms have leftovers in low demand and do not reorder in high demand.

3. For $x_i = x_{-i} \in (HB(\alpha_L), OU(\alpha_H))$, the derivative of firm i's expected profit is constant. (i) If $c \ne C/2$, the derivative is nonzero, so no point in $(HB(\alpha_L), OU(\alpha_H))$ can be a symmetric equilibrium because an equilibrium has to be a local maximizer. (ii) If $c = C/2$, then the derivative is zero, so every point in $(HB(\alpha_L), OU(\alpha_H))$ is an equilibrium candidate. However, no such point can be a Pareto-dominant equilibrium. This follows because $c = C/2$ implies $r_{c_2} < 1/2 = r_c$ which implies by point 1. above that x_l^{R*} is a symmetric equilibrium. Noting that $x_l^{R*} < HB(\alpha_L)$, it follows that the expected profit for $x_i = x_{-i} = x_l^{R*}$ strictly exceeds that for $x_i = x_{-i} \in [HB(\alpha_L), OU(\alpha_H)]$.

4. Let $x_o^{R*} := OU(\alpha_H)$. The point $x_i = x_{-i} = x_o^{R*}$ is a symmetric equilibrium candidate if and only if $c \le C/2$ and $r_c \ge r_{c_1}$. This holds because for $x_{-i} = OU(\alpha_H)$, $\Pi_i^R(\mathbf{x})$ is not differentiable at $x_i = OU(\alpha_H)$ which is a local maximizer if the left derivative

$C/2 - c \geq 0$ and the right derivative $\frac{1}{2}\left(\frac{\alpha_H}{1-\gamma} - x_i\frac{2+\gamma}{1-\gamma^2}\right)c$ is nonpositive, which holds if and only if $r_c \geq r_{c_1}$. Furthermore, we can show that x_o^{R*} is a symmetric equilibrium if $r_c \in [r_{c_1}, r_{c_2}]$ when $r_\alpha \geq 2$. Under this equilibrium, firms have leftovers in low demand and reorder in high demand. Finally, x_o^{R*} is not a Pareto-dominant equilibrium if $r_c \geq r_{c_2}$: In this case x_l^{R*} is a symmetric equilibrium by point 1. above, and it is straightforward to verify that the resulting expected profit exceeds that for $x_i = x_{-i} = x_o^{R*}$.

Table 2.4: Sufficient and Necessary Conditions for Equilibria ($1 \leq r_\alpha < 2$)

Equilibrium Candidate ($1 \leq r_\alpha < 2$)	Conditions	Strategy in Low Demand	Strategy in High Demand
$x_h^{R*} := \frac{1-\gamma^2}{2+\gamma}C(r_\alpha - 2r_c) > HB(\alpha_L)$	$0 \leq r_c < r_{c_4}$ or $\max\{r_{c_4}, r_{c_5}\} \leq r_c \leq r_{c_7}$	leftover	no reorder
$x_p^{R*} := \frac{1-\gamma^2}{2(2+\gamma)}C(r_\alpha - 2r_c + 1) \in (OU(\alpha_H), HB(\alpha_L))$	$\max\{0, r_{c_4}\} \leq r_c \leq r_{c_5}$ and $0 \leq r_\alpha \leq \hat{r}\alpha$	clear	no reorder
$x_o^{R*} := \frac{1-\gamma}{2-\gamma}C(r_\alpha - 1) = OU(\alpha_H)$	$\max\{r_{c_5}, r_{c_7}\} < r_c < r_{c_6}$	clear	no reorder
$x_l^{R*} := \frac{2(1-\gamma^2)}{2+\gamma}C(1 - r_c) \leq OU(\alpha_H)$	$r_{c_6} \leq r_c \leq 1$	clear	reorder

Second, we consider $1 < r_\alpha < 2$, i.e., $\alpha_L \leq \alpha_H < 2\alpha_L$. Then $0 = OU(\alpha_L) < OU(\alpha_H) < HB(\alpha_L) < HB(\alpha_H)$. Following a similar procedure, we show that there are four equilibrium candidates for $r_\alpha \in (1, 2)$. Table 2.4 summarizes the sufficient and necessary equilibrium conditions for each candidate, where $\hat{r}\alpha$ is the r_α-axis coordinate of the intersection of the functions r_{c_4} and r_{c_5}.

The above arguments prove the existence of a symmetric equilibrium. The equilibrium strategies for the cases (i)-(iii) in the statement of Proposition 2.3 are obtained by summarizing points 1-4 for $r_\alpha \geq 2$ and Table 2.4 for $1 < r_\alpha < 2$, letting $r_c(r_\alpha, \gamma) = r_{c_2}\mathbf{1}_{\{r_\alpha \geq 2\}}$, $\overline{r}_c(r_\alpha, \gamma) = r_{c_6}\mathbf{1}_{\{1 \leq r_\alpha < 2\}}$, and $\underline{r}_c(r_\alpha, \gamma) = \left(r_{c_4}\mathbf{1}_{\{r_{c_5} \geq r_{c_4}\}} + r_{c_7}\mathbf{1}_{\{r_{c_5} \leq r_{c_7}\}}\right)\mathbf{1}_{\{1 \leq r_\alpha < 2\}}$. (Note that $r_{c_4} \leq 0$ for $r_\alpha \leq m^{**}(\gamma)$, in which case the conditions on r_c in the first two rows in Table 2.4 simplify.) ■

Proof of Proposition 2.4. We first prove the results for $r_\alpha \geq 2$ and then extend to $1 \leq r_\alpha < 2$.

Table 2.5 shows that there are 6 possible no-reorder/reorder equilibrium combinations for any given r_c and r_α such that $r_\alpha \geq 2$. We need to quantify the expected profit differences for all equilibrium combinations. By Proposition 2.2, we know that there is at least one symmetric equilibrium in the N game for a given parameter pair (r_α, r_c). If $\alpha_L = C(1-\gamma)$ and $\alpha_H \geq 2\alpha_L$, we assume that both firms adopt the Pareto-dominant x_l^{N*} over x_h^{N*} when two symmetric equilibria exist. Propositions 2.2 and 2.3 give us the initial equilibrium ordering quantities. Referring to Table 2.2, we can also calculate

the corresponding equilibrium prices and obtain the expected equilibrium profits. Table 2.6 lists the conditional revenue and cost in low and high scenarios for all equilibria. Note that conditions 1.(a)-(c) hold whenever the firms with reorder flexibility are more profitable than inflexible firms. We show that 1.(a)-(c) imply $x^{R*} < x^{N*} < OU\,(\alpha_H)$.

Table 2.5: Equilibrium Pairs

	$r_\alpha \geq 2$			$1 \leq r_\alpha < 2$			
	$r_c \in [0, r_{c_1})$	$[r_{c_1}, r_{c_3}]$	$[r_{c_2}, 1]$	$r_c \in [0, r_{c_4}]$	$[0, r_{c_5}]$	(r_{c_5}, r_{c_6})	$[r_{c_6}, 1]$
$r_c \geq \frac{r_\alpha}{2} - \frac{m^{**}(\gamma)}{2}$	(x_l^{N*}, x_h^{R*})	(x_l^{N*}, x_o^{R*})	(x_l^{N*}, x_l^{R*})	(x_l^{N*}, x_h^{R*})	(x_l^{N*}, x_p^{R*})	(x_l^{N*}, x_o^{R*})	(x_l^{N*}, x_l^{R*})
$r_c < \frac{r_\alpha}{2} - \frac{m^{**}(\gamma)}{2}$	(x_h^{N*}, x_h^{R*})	(x_h^{N*}, x_o^{R*})	(x_h^{N*}, x_l^{R*})	(x_h^{N*}, x_h^{R*})	(x_h^{N*}, x_p^{R*})	(x_h^{N*}, x_o^{R*})	N/A

We compare the expected profits of each equilibrium combination as follows.

(i) $\{(r_\alpha, r_c) \mid r_c < r_\alpha/2 - m^{**}(\gamma)/2 \text{ and } r_{c_2} \leq r_c \leq 1\}$. In this case, firms choose x_h^{N*} in the N game and x_l^{R*} in the R game. By Table 2.6, we have x_l^{R*} generates more expected profit than x_l^{N*} if and only if $(r_\alpha, r_c) \in F_1$, where

$$F_1 := \left\{(r_\alpha, r_c) \mid r_\alpha > 2 \text{ and } \frac{\gamma^3}{2(1+\gamma)(2-\gamma)^2} r_\alpha + \frac{1}{2} < r_c < r_\alpha/2 - m^{**}(\gamma)/2 \leq 1\right\}.$$

The set F_1 is nonempty if and only if $2 + m^{**}(\gamma) \geq \frac{(\gamma+1)(2-\gamma)^2}{\gamma^3}$, i.e, $\gamma \leq 0.875$. Note that for nonempty F_1, we have $r_c > r_{c_3}$, which implies that x_l^{R*} is the unique symmetric R game equilibrium candidate.

(ii) $\{(r_\alpha, r_c) \mid r_c \geq r_\alpha/2 - m^{**}(\gamma)/2 \text{ and } r_{c_2} \leq r_c \leq 1\}$. In this case, firms choose x_l^{N*} in the N game and x_l^{R*} in the R game. By Table 2.6, we have x_l^{N*} generates more profit than x_l^{R*} if and only if $(r_\alpha, r_c) \in F_2 := \{(r_\alpha, r_c) \mid 2 \leq r_\alpha \leq 2r_c + m^{**}(\gamma) \text{ and } 1 - \frac{1}{2}(r_\alpha - 1)(1 - 2\frac{\gamma\sqrt{\gamma(\gamma+1)}}{(\gamma+1)(2-\gamma)}) < r_c \leq 1\}$. The set F_2 is nonempty if and only if $\gamma \leq \frac{1}{3}\left((17 + 12\sqrt{2})^{1/3} + (17 + 12\sqrt{2})^{-1/3} - 1\right) \approx 0.849$. Note that for nonempty F_2, we have $r_c > r_{c_3}$, which implies that x_l^{R*} is the unique symmetric R game equilibrium.

(iii) $\{(r_\alpha, r_c) \mid r_c < r_\alpha/2 - m^{**}(\gamma)/2 \text{ and } r_{c_1} \leq r_c \leq r_{c_3}\}$. In this case, although x_l^{R*} may also be a reorder symmetric equilibrium, the analysis is similar to case (i). Thus here we only compare the equilibrium combination x_h^{N*} and x_o^{R*}. Algebraically, x_o^{R*} generates more profit than x_h^{N*} if and only if

$$(r_\alpha, r_c) \in \left\{(r_\alpha, r_c) \mid r_\alpha \geq 2 \text{ and } \frac{\gamma - 2}{\gamma} r_c + \frac{2+\gamma}{2\gamma} \leq r_\alpha \leq \frac{2(\gamma+1)(\gamma-2)}{\gamma^2} r_c + \frac{2+\gamma}{\gamma^2}\right\},$$

which is exclusive of the validity set $\{(r_\alpha, r_c) \mid r_c < r_\alpha/2 - m^{**}(\gamma)/2 \text{ and } r_{c_1} \leq r_c \leq r_{c_3}\}$. Thus, we conclude that the expected profit of x_o^{R*} is no more than that of x_h^{N*} in this

Table 2.6: Revenue and Cost at Equilibrium for $\alpha_H \geq 2\alpha_L$

	Equilibrium	Conditional on the Low Scenario		Conditional on the High Scenario	
		Revenue	Cost	Revenue	Cost
R game	x_l^{R*}	$\left(\frac{\alpha_L}{1-\gamma} - \frac{1}{1-\gamma}x_l^{R*}\right)x_l^{R*}$	cx_l^{R*}	$\frac{\alpha_H+C}{2-\gamma}OU(\alpha_H)$	$cx_l^{R*} + C(OU(\alpha_H) - x_l^{R*})$
	x_o^{R*}	$\frac{\alpha_L}{2-\gamma}HB(\alpha_L)$	cx_o^{R*}	$\frac{\alpha_H+C}{2-\gamma}OU(\alpha_H)$	cx_o^{R*}
	x_h^{R*}	$\frac{\alpha_L}{2-\gamma}HB(\alpha_L)$	cx_h^{R*}	$\left(\frac{\alpha_H}{1-\gamma} - \frac{1}{1-\gamma}x_h^{R*}\right)x_h^{R*}$	cx_h^{R*}
N game	x_l^{N*}	$\left(\frac{\alpha_L}{1-\gamma} - \frac{1}{1-\gamma}x_l^{N*}\right)x_l^{N*}$	cx_l^{N*}	$\left(\frac{\alpha_H}{1-\gamma} - \frac{1}{1-\gamma}x_l^{N*}\right)x_l^{N*}$	cx_l^{N*}
	x_h^{N*}	$\frac{\alpha_L}{2-\gamma}HB(\alpha_L)$	cx_h^{N*}	$\left(\frac{\alpha_H}{1-\gamma} - \frac{1}{1-\gamma}x_h^{N*}\right)x_h^{N*}$	cx_h^{N*}

case.

(iv) $\{(r_\alpha, r_c) \mid r_c \geq r_\alpha/2 - m^{**}(\gamma)/2 \text{ and } r_{c_1} \leq r_c \leq r_{c_3}\}$. In this case, although x_l^{R*} may also be a reorder symmetric equilibrium, the analysis is similar to case (ii). We here only compare the equilibrium combination x_l^{N*} and x_o^{R*}. By Table 2.6, we calculate the profit difference $\Pi_i^N(x_i = x_{-i} = x_l^{N*}) - \Pi_i^R(x_i = x_{-i} = x_o^{R*})$

$$= \left(\frac{1-\gamma^2}{4(2+\gamma)^2} - \frac{1}{2}\left(\frac{1-\gamma}{2-\gamma}\right)^2\right)r_\alpha^2 - \left(\frac{1-\gamma^2}{(2+\gamma)^2} - \frac{1-\gamma}{2-\gamma}\right)r_\alpha r_c + \frac{1-\gamma^2}{(2+\gamma)^2}r_c^2$$
$$+ \left(\frac{1-\gamma^2}{2(2+\gamma)^2} + \left(\frac{1-\gamma}{2-\gamma}\right)^2 - \frac{1-\gamma}{2(2-\gamma)}\right)r_\alpha + \left(-\frac{1-\gamma^2}{(2+\gamma)^2} - \frac{1-\gamma}{2-\gamma}\right)r_c + \frac{1-\gamma^2}{4(2+\gamma)^2} - \frac{1}{2}\frac{1-\gamma}{(2+\gamma)^2}.$$

Now we prove that this lower bound is nonnegative. Substituting $r_c = r_\alpha/2 - m^{**}(\gamma)/2$ into $\Pi_i^N(x_i = x_{-i} = x_l^{N*}) - \Pi_i^R(x_i = x_{-i} = x_o^{R*})$ and noticing $0 \leq \gamma \leq 1$, we can show the profit difference is nonnegative.

(v) $\{(r_\alpha, r_c) \mid r_c < r_\alpha/2 - m^{**}(\gamma)/2 \text{ and } 0 \leq r_c < r_{c_1}\}$. In this case, we compare the equilibrium combination x_h^{N*} and x_h^{R*}. By Propositions 3 and 4, $x_h^{R*} = x_h^{N*}$.

(vi) $\{(r_\alpha, r_c) \mid r_c \geq r_\alpha/2 - m^{**}(\gamma)/2 \text{ and } 0 \leq r_c < r_{c_1}\}$. In this case, we compare the equilibrium combination x_l^{N*} and x_h^{R*}. Note that the profit of x_h^{R*} has an identical expression as that of x_h^{N*} and we can verify that if $\alpha_L = C(1-\gamma)$ and $\alpha_H \geq 2\alpha_L$, then $\Pi_i^N(x_i = x_{-i} = x_l^{N*}) \geq \Pi_i^N(x_i = x_{-i} = x_h^{N*})$ if and only if

$$(r_\alpha, r_c) \in \left\{(r_\alpha, r_c) \mid \frac{1}{2}\left(r_\alpha - 1 - \frac{2\gamma}{2-\gamma}\sqrt{\frac{\gamma}{1+\gamma}}\right) \leq r_c \leq \frac{1}{2}\left(r_\alpha - 1 + \frac{2\gamma}{2-\gamma}\sqrt{\frac{\gamma}{1+\gamma}}\right)\right\},$$

which contains the set of $\{(r_\alpha, r_c) \mid r_c \geq r_\alpha/2 - m^{**}(\gamma)/2 \text{ and } 0 \leq r_c < r_{c_1}\}$. Thus in this case, x_l^{N*} generates more profit than x_h^{R*}.

From (i) to (vi), we see that reorder flexibility benefits firms only in (i) and (ii), where firms play only x_l^{R*} in the R game but have different no-reorder equilibrium quantity x_l^{N*} and x_h^{N*} dependent on (r_α, r_c). Moreover, $x_l^{R*} \leq x_l^{N*}$ if and only if $(r_\alpha, r_c) \in$

$\{(r_\alpha, r_c) \mid r_c \geq -r_\alpha/2 + 3/2\} \supset F_1 \cup F_2$. Thus, we conclude that $x_l^{R*} < x_l^{N*} < OU(\alpha_H) < x_h^{N*}$. The parameter subset where reorder flexibility benefits firms is $F_1 \cup F_2$. This set is nonempty if $\gamma \leq 0.849$.

Second, let us consider the case where $1 \leq r_\alpha < 2$. (i) Whenever x_h^{R*} and x_p^{R*} are equilibria, the equilibrium outcomes of the R game are respectively equivalent to the outcome of x_h^{N*} and x_l^{N*} in the N game. Thus, reorder flexibility has no value when the equilibrium is either x_h^{R*} or x_p^{R*}. (ii) We explore the value of reorder flexibility when x_o^{R*} is an equilibrium. The algebraic expression $\Pi_i^N(x_i = x_{-i} = x_l^{N*}) - \Pi_i^R(x_i = x_{-i} = x_o^{R*}) > 0$ if and only if $(r_\alpha, r_c) \in \left\{(r_\alpha, r_c) \mid r_\alpha \geq 1 \text{ and } 1 - \frac{2+3\gamma}{2(2-\gamma)}(r_\alpha - 1) < r_c < 1 - \frac{\gamma^2+\gamma+2}{2(2-\gamma)(1+\gamma)}(r_\alpha - 1)\right\}$. However, in this region x_o^{R*} is not an equilibrium. Thus, reorder flexibility has no positive value when x_o^{R*} is an equilibrium. (iii) Consider x_l^{R*}. From case (ii) of $r_\alpha \geq 2$, we have that reorder flexibility has a positive value for $(r_\alpha, r_c) \in F_3$, where

$$F_3 := \left\{(r_\alpha, r_c) \mid 1 \leq r_\alpha < 2 \text{ and } 1 - \frac{1}{2}(r_\alpha - 1)\left(1 - 2\frac{\gamma\sqrt{\gamma(\gamma+1)}}{(\gamma+1)(2-\gamma)}\right) < r_c \leq 1\right\}.$$

In sum, the reorder flexibility has a positive value only when x_l^{R*} is an equilibrium and $(r_\alpha, r_c) \in F_1 \cup F_2 \cup F_3$. Let $\underline{r}\alpha(\gamma) := 1 + \left(\left(\frac{\gamma^3}{4-3\gamma^2} - 1\right)m^{**}(\gamma) + \left(\frac{\gamma^3}{4-3\gamma^2} + 1\right)\right)\mathbf{1}_{\{0.849 \leq \gamma < 0.875\}}$, $\overline{r}\alpha(\gamma) := (1+\gamma)(2-\gamma)^2/\gamma^3$ and

$$\underline{r}_c(\gamma, r_\alpha) := \begin{cases} 1 + \left(\frac{\gamma}{2-\gamma}\sqrt{\frac{\gamma}{1+\gamma}} - \frac{1}{2}\right)(r_\alpha - 1) & \text{if } \underline{r}\alpha(\gamma) \leq r_\alpha < \tilde{r}\alpha(\gamma), \\ \frac{\gamma^3}{2(1+\gamma)(2-\gamma)^2}r_\alpha + \frac{1}{2} & \text{if } \tilde{r}\alpha(\gamma) \leq r_\alpha < \overline{r}\alpha(\gamma), \end{cases}$$

where $\tilde{r}\alpha(\gamma) = 1 + (m^{**}(\gamma) + 1)/\left(2 - \frac{\gamma}{2-\gamma}\sqrt{\frac{\gamma}{1+\gamma}}\right)$. Moreover, the R game has the same equilibrium as the N game if and only if $r_\alpha \leq \overline{\overline{r}}_\alpha(\gamma) = \frac{2+\gamma}{\gamma^2}$ and $r_c \leq \overline{r}_c(\gamma, r_\alpha) = \max\{r_{c_4}, r_{c_5}\}\mathbf{1}_{\{1 \leq r_\alpha < 2\}} + r_{c_1}\mathbf{1}_{\{r_\alpha \geq 2\}}$. ■

Proof of Proposition 2.5. We first solve the 2nd-stage pricing-ordering game given the 1st-stage orders, and then determine the equilibrium orders in the 1st stage.

Define $\hat{x}_I = \frac{(2+\gamma)\alpha + \gamma C}{4-\gamma^2}$ and $\hat{x}_F = \frac{(2+\gamma)\alpha - (2-\gamma^2)C}{4-\gamma^2}$. Following the proof of Lemma 2.1, we can show that for any initial inventory vector $\mathbf{x} = (x_I, x_F)$ where x_I and x_F represent inflexible and flexible firms' inventory, the price subgame of the U game has a unique equilibrium as follows:

- for the inflexible firm, it prices to clear its inventory if $x_I \leq \overline{x}_I(x_F)$ and prices to have leftover otherwise;
- for the flexible firm, (i) if $x_F < \underline{x}_F(x_I)$ then it reorders and prices to clear its inventory; (ii) if $\underline{x}_F(x_I) \leq x_F \leq \overline{x}_F(x_I)$ then it prices to clear its inventory but does not reorder;

Table 2.7: Notation in Proof of Proposition 2.5

$y_1 = 1/(2\rho_9\gamma(1,-10,26,4,-100,88,104,-144,0,64))$
$y_2 = \rho_{11}\gamma(-2,3,21,-20,-101,43,252,0,-304,$
$\quad -96,128,64)$, $y_3' = (\gamma^2-\gamma-2)^2$
$y_3 = \rho_7\gamma(5,12,-34,-74,68,136,-32,-64)y_3'$
$y_4 = \rho_{11}\gamma(2,16,-56,-126,337,441,-816,$
$\quad -848,832,816,-256,-256)$
$y_5 = \rho_9\gamma(1,-5,-4,28,32,-100,-64,144,32,-64)$
$y_6 = \rho_8\gamma(-9,54,-6^2,-236,284,312,-432,-2^5,192)$
$y_7 = 4\rho_{14}\gamma(0.5,0,-33,96,136,-656,-388,1808,$
$\quad 1048,-2528,-1792,,1792,1536,-512,-512)$
$y_8 = 2\sqrt{2}\gamma^3(\gamma^2+\gamma-2)(\gamma^2-2)(\gamma^4-14\gamma^2+16)$
$y_9 = (\gamma-1)(\gamma+2)\rho_{12}\gamma(1,0,-40,96,280,-464,$
$\quad -904,880,1568,-768,-1408,256,512)$
$y_{10} = -2^{\frac{3}{2}}\gamma^2(\gamma^2-2)(\gamma^4-14\gamma^2+16)\rho_3\gamma(2,-1,-4,4)$
$y_{11} = \rho_{13}\gamma(-1,-48,280,-464,-1416,3152,3144,$
$\quad -7056,4064,7744,3072,-4352,-1024,1024)$
$r_{c_8} = (8(1-\gamma)^{\frac{3}{2}}(y_2r_\alpha^2+y_3r_\alpha+y_4)^{\frac{1}{2}}+y_5r_\alpha+y_6)y_1$
$r_{c_9} = ((3\gamma^4-7\gamma^2+4)^{\frac{1}{2}}(y_8r_\alpha+y_{10})+y_9r_\alpha+y_{11})/y_7$
$r_{c_{10}} = (-2\gamma^2-\gamma+4)(2(2-\gamma^2))^{-1}$,
$r_{c_{11}} = (-\gamma(\gamma+2)r_\alpha+\rho_3\gamma(1,-13,-2,16))\times$
$\quad (\rho_3\gamma(2,-12,-4,16))^{-1}$
$r_{c_{12}} = (1/\rho_4\gamma(1,1,-4,-2,4))\left((2\gamma^4-8\gamma^2+8)r_\alpha+\right.$
$\quad \left.\rho_4\gamma(2,1,-6,-2,4)+\gamma^2\sqrt{2(3\gamma^4-7\gamma^2+4)}\ \right)$

(iii) if $x_F > \overline{x}_F(x_I)$ then it prices to have leftover and does not reorder;

where

$$\underline{x}_F(x_I) = \begin{cases} -\frac{\gamma}{2-\gamma^2}x_F + \frac{(1+\gamma)(\alpha-C+\gamma C)}{2-\gamma^2} & \text{if } x_I < \hat{x}_I \\ \hat{x}_F, & \text{if } x_I \geq \hat{x}_I \end{cases}, \overline{x}_F(x_I) = \begin{cases} -\frac{\gamma}{2-\gamma^2}x_F + \frac{(1+\gamma)\alpha}{2-\gamma^2} & \text{if } x_I < HB(\alpha) \\ HB(\alpha) & \text{if } x_I \geq HB(\alpha) \end{cases},$$

and

$$\overline{x}_I(x_F) = \begin{cases} \hat{x}_I & \text{if } x_F \leq \hat{x}_F \\ -\frac{\gamma}{2-\gamma^2}x_F + \frac{(1+\gamma)\alpha}{2-\gamma^2} & \text{if } \hat{x}_F < x_F < HB(\alpha). \\ HB(\alpha) & \text{if } x_F \geq HB(\alpha) \end{cases}$$

Solving the 1st stage ordering game proceeds similarly as in the proofs of Propositions 2.2 and 2.3. We first identify all possible equilibrium candidates and then characterize the conditions which ensure that a candidate is the best response. As reorder flexibility is asymmetric in the U game, we need to establish these conditions both for the flexible and for the inflexible firm. As the idea of the proof is essentially the same as Propositions 2.2 and 2.3, we omit the details of the rather complicated algebra and directly present the results. Let us define $\check{r}\alpha$ as the r_α-axis coordinate of the intersection of r_{c_4} and r_{c_7}. Then the following holds for the three cases:

(*i*) If $r_c \leq \underline{r}_c(r_\alpha,\gamma) := r_{c_5}\mathbf{1}_{\{1\leq r_\alpha<\check{r}\alpha\}} + r_{c_7}\mathbf{1}_{\{\check{r}\alpha\leq r_\alpha<2\}} + r_{c_1}\mathbf{1}_{\{r_\alpha\geq 2\}}$, then the U game has

the same initial order equilibrium as the R and N games. In particular,

$$\mathbf{x}^{U*} = \begin{cases} \mathbf{x}_h^{R*} = \mathbf{x}_h^{N*}, & \text{if } (r_\alpha, r_c) \in \left\{(r_\alpha, r_c) | 0 \le r_c < r_{c_1}\mathbf{1}_{\{r_\alpha \ge 2\}} + r_{c_4}\mathbf{1}_{\{1 \le r_\alpha < 2\}}\right\} \\ & \qquad \cup \left\{(r_\alpha, r_c) | \check{r}\alpha \le r_\alpha < 2 \text{ and } \max\{r_{c_4}, r_{c_5}\} \le r_c < r_{c_7}\right\}; \\ \mathbf{x}_p^{R*} = \mathbf{x}_l^{N*}, & \text{if } (r_\alpha, r_c) \in \left\{(r_\alpha, r_c) | 1 \le r_\alpha < \hat{r}\alpha \text{ and } \max\{0, r_{c_4}\} \le r_c \le r_{r_5}\right\}. \end{cases}$$

(ii) If $\underset{\sim}{r}_c(r_\alpha, \gamma) < r_c < \tilde{r}_c(r_\alpha, \gamma)$, there is a continuum of equilibria on the line $y = -\frac{\gamma}{2-\gamma^2}x + \frac{(1-\gamma^2)(r_\alpha - 1)C}{2-\gamma^2}$ in the one-sided ϵ-neighbourhood $(x_l - \epsilon, x_l]$ of x where $x_l = \frac{C}{4}\left(-2(2-\gamma^2)r_c + (\gamma+2)(1-\gamma) + \gamma\right)$.

(iii) If $\tilde{r}_c(r_\alpha, \gamma) := r_{c_6}\mathbf{1}_{\{1 \le r_\alpha < 2\}} + \max\{r_{c_8}, 1/2\}\mathbf{1}_{\{r_\alpha \ge 2\}} \le r_c \le 1$, then the U game has a unique asymmetric initial order equilibrium. At equilibrium, the flexible firm reorders if the market is realized as high. In particular[5],

$$\mathbf{x}^{U*} = \begin{cases} ((x_I^{U*})_{cn}, (x_F^{U*})_{cn}), & \text{if } (r_\alpha, r_c) \in \left\{(r_\alpha, r_c) | \max\{r_{c_9}, \max\{r_{c_8}, 1/2\}\}\mathbf{1}_{\{r_\alpha \ge 2\}}\right. \\ & \qquad \left. + r_{c_6}\mathbf{1}_{\{1 \le r_\alpha < 2\}} \le r_c < r_{c_{12}}\right\}; \\ ((x_I^{U*})_{ln}, (x_F^{U*})_{ln}), & \text{if } (r_\alpha, r_c) \in \{(r_\alpha, r_c) | 1/2 \le r_c < \max\{r_{c_9}, r_{c_{10}}\}\}; \\ ((x_I^{U*})_{cz}, (x_F^{U*})_{cz}), & \text{if } (r_\alpha, r_c) \in \{(r_\alpha, r_c) | \max\{r_{c_{11}}, r_{c_{12}}\} \le r_c \le 1\}; \\ ((x_I^{U*})_{lz}, (x_F^{U*})_{lz}), & \text{otherwise;} \end{cases}$$

where

$$\begin{aligned} (x_I^{A*})_{cn} &= \tfrac{2(1-\gamma^2)C}{\gamma^4 - 14\gamma^2 + 16}\left(\gamma^3 - \gamma^2 - \gamma + 2 + (-\gamma^3 + 2\gamma^2 + 2\gamma - 4)r_c + (\gamma+2)(1-\gamma)r_\alpha\right); \\ (x_F^{U*})_{cn} &= \tfrac{(1-\gamma^2)C}{\gamma^4 - 14\gamma^2 + 16}\left(\gamma^3 - 13\gamma^2 - 2\gamma + 16 - (2\gamma^3 - 12\gamma^2 - 4\gamma + 16)r_c - (\gamma^2 + 2\gamma)(1-\gamma)r_\alpha\right); \\ (x_I^{U*})_{cz} &= \tfrac{(1-\gamma^2)C}{2(4-3\gamma^2)}\left((1+\gamma)(2-\gamma) - 2(2-\gamma^2)r_c + (\gamma+2)(1-\gamma)r_\alpha\right); \\ (x_F^{U*})_{lz} &= \tfrac{C}{4}\left(-2\gamma^2 - \gamma + 4 + 2(\gamma^2 - 2)r_c\right); \quad (x_F^{U*})_{cz} = (x_F^{U*})_{lz} = 0; \\ (x_I^{U*})_{ln} &= (x_I^{U*})_{lz} = \tfrac{C}{4}\left(\gamma + 2(\gamma^2 - 2)r_c + (\gamma+2)(1-\gamma)r_\alpha\right). \end{aligned}$$

The proofs of $x_F^{U*} > x^{N*}$ for case *(ii)* and $x_F^{U*} < x^{R*} < x^{N*}$ for case *(iii)* involve lengthy but straightforward algebra and are therefore omitted. ■

Proof of Proposition 2.6. For a given (r_α, r_c) in this region, the N, R, and U games each have a unique equilibrium. The proof proceeds by showing at least one of the inequalities $\Pi_F^{U*} < \Pi^{N*}$ and $\Pi_I^{U*} < \Pi^{R*}$ holds. Table 2.9 summarizes under what condition which firm would deviate from the asymmetric flexibility strategies (N, R). ■

[5]The first subscript c or l represents the inflexible firm's strategy, c for *clear-out* and l for *left-over*. The second subscript represents the flexible firm's strategy, n for *non-zero* and z for *zero*.

Table 2.8: Notation in Proof of Proposition 2.6

$z_1 = \rho_{11}\gamma(1, 4, -44, -192, 168, 1280, 344, -3008, -2000, 2944, 2560, -1024, -1024)$

$z_2 = \rho_{12}\gamma(-3, -6, 58, 188, -116, -1104, -528, 2592, 2128, -2688, -2560, 1024, 1024)$

$z_3 = \rho_7\gamma(-1, -2, 16, 32, -44, -88, 32, 64)$

$z_4 = \rho_7\gamma(4, -8, -9, 64, 6, -120, 0, 64)$

$z_5 = \left((1-\gamma)(\gamma-2)(\gamma^4-\gamma^3+16\gamma+16)\gamma^3\right)^{-1}$

$z_6 = (1-\gamma)(2-\gamma)(2-\gamma^2)\rho_5\gamma(1, 2, 8, -8, -48, -32)$

$z_7 = (1-\gamma)(2-\gamma)(3\gamma^2-4)(\gamma^2-2\gamma-4)(\gamma+2)^2$

$z_8 = 4(1-\gamma^2)(4-\gamma^2)^2(3\gamma^2-4)^2(2-\gamma^2)^2$

$z_9 = (\gamma+2)^2(2-\gamma^2)^2\rho_{10}\gamma(1, -4, -3, 48, -34, -132, 124, 144, -144, -64, 64)$

$z_{10} = \left(2\gamma^4(\gamma^2-2)(\gamma+2)^2(1-\gamma)^2\right)^{-1}$

$z_{11} = 4(\gamma-1)(\gamma-2)^2(2-\gamma^2)^2(\gamma+2)^3$

$z_{12} = 2\rho_9\gamma(1, 9, 4, -66, -44, 192, 96, -2^8, -2^6, 2^7)$

$z_{13} = 8(1-\gamma)^2(4-\gamma^2)^2(\gamma^2-2)^3\rho_8\gamma(1, 4, 4, -64, -16, 256, 192, -256, -256)$

$z_{14} = 8(1-\gamma)^2(4-\gamma^2)^2(\gamma^2-2)^2\rho_{10}\gamma(2, 5, 8, -88, -64, 400, 256, -768, -640, 512, 512)$

$z_{15} = \rho_{20}\gamma(8, -8, -70, -148, 310, 3568, -3180, -26448, 26608, 102144, -116096, -239104, 289280, 364544, -442368, -368640, 425984, 229376, -245760, -65536, 65536)$

$\chi_{cn} = 1 + 2z_1(r_c-1)/(z_2 - \gamma z_3\sqrt{2\gamma z_4})$

$\chi_{lz} = z_5\left(z_6 r_c + z_7 - 4\sqrt{(r_\alpha^2 - r_\alpha)z_8 + z_9}\right)$

$\chi_{ln} = z_{10}\left(z_{11}r_c + z_{12} - \sqrt{z_{13}r_c^2 - z_{14}r_c + z_{15}}\right)$

Table 2.9: Deviating Firm in Asymmetric Reorder Flexibility Endowment (N, R)

Equil.	Condition	Deviating Firm	Equil.	Condition	Deviating Firm
$\mathbf{x}_{cz}^{U*}$	any (r_α, r_c)	inflexible $(\Pi_I^{U*} < \Pi^{R*})$	$\mathbf{x}_{lz}^{U*}$	$r_\alpha < \chi_{lz}(r_c, \lambda)$	inflexible $(\Pi_I^{U*} < \Pi^{R*})$
				$r_\alpha > \chi_{lz}(r_c, \lambda)$	flexible $(\Pi_F^{U*} < \Pi^{N*})$
$\mathbf{x}_{cn}^{U*}$	$r_\alpha < \chi_{cn}(r_c, \lambda)$	inflexible $(\Pi_I^{U*} < \Pi^{R*})$	$\mathbf{x}_{ln}^{U*}$	$r_\alpha < \chi_{ln}(r_c, \lambda)$	inflexible $(\Pi_I^{U*} < \Pi^{R*})$
	$r_\alpha > \chi_{cn}(r_c, \lambda)$	flexible $(\Pi_F^{U*} < \Pi^{N*})$		$r_\alpha > \chi_{ln}(r_c, \lambda)$	flexible $(\Pi_F^{U*} < \Pi^{N*})$

Chapter 3

Efficient Information Heterogeneity in a Queue

3.1 Introduction

In today's service industries, information about delays is ubiquitous. The border-crossing waiting time between US and Canada is posted online and updated in almost real time. Information about traffic jams on major roads is distributed in real time on radio, television and Internet. Thanks to traffic-information sharing apps like Waze, real-time information about traffic may also be available even for roads that are not covered by governmental-funded traffic detection and monitoring.

Regardless of how widely available information about real-time delays may be, a large fraction of customers are still uninformed. First, not all people are equipped with mobile devices that may make information acquisition almost effortless. Second, people may simply overlook up-to-minute information about delays before hitting the road. In an online poll with about 20,000 participants who were asked, "How do you most often check traffic information before going out?", 47% answered "I don't check"; the rest checked various sources such as TV, radio, computer, and mobile devices.[1] It may be hard to ascertain the exact causes of information ignorance, which may be numerous. For example, some people may simply be over-confident in their luck. Such information ignorance may also manifest itself as that people sometimes check information and sometimes do not. Last, there could be other reasons leading to information heterogeneity. For example, small service providers may not afford to invest in technology for tracking and reporting how crowded they are. In this case, only drop-in customers can observe the queue, and

[1]The poll result can be found at `http://www.gasbuddy.com/GB_Past_Polls.aspx?poll_id=720`.

many potential customers are not.

Thus it is evident that many of today's service environments are characterized by heterogeneity in their customers' knowledge about delays: some are informed about the real-time delay, whereas others are not but have had past experiences with the congestion levels. It is essential to understand the interaction among customers with information heterogeneity in order to answer the question: How do system throughput and social welfare change, as the real-time delay information becomes more prevalent due to advances in information technology?

On the one hand, Chen and Frank (2004) show that delay information is a double-edged sword for system throughput by comparing full and no real-time delay information. When the system load is low, customers might be turned away with real-time information, but otherwise might stay if uninformed. Hence, the throughput of an observable queueing system only exceeds that of the unobservable counterpart when the system load is high. In a more common situation where some, but not all, customers are informed, would the system throughput outperform its counterparts in the two extreme information structures, i.e., full and no information?

On the other hand, Hassin (1986) argues that real-time congestion information can effectively improve social welfare, again by comparing full and no real-time delay information. The intuition is that it helps better match capacity with customer demand intertemporally: customers never join a long queue or balk from a short one. This rationale is consistent with the ubiquity of congestion information in today's public service industries. However, as we argued, it might be too ideal to expect that all customers have access to delay information even if it is readily available. More essentially, does the system inevitably suffer efficiency loss due to the presence of uninformed customers?

Service congestion is often modeled and studied by applying queueing theory. In that literature, the comparison between observable and unobservable queues has been well studied. For examples, an influential work Hassin (1986) and a follow-up paper Chen and Frank (2004) consider a single-server queue where customers arrive according to a Poisson process and service takes an exponential time. The authors of the latter paper show that there exists a critical level of the implied utilization (i.e., the potential arrival rate divided by the service rate) such that, if the implied utilization is beyond the critical level (which can be negative), the system throughput is higher in the observable setting than in the unobservable setting; and if the implied utilization is lower than the critical level, the reverse is the case[2]. However, it is unclear how one can apply the results obtained by comparing two extreme scenarios of information homogeneity – full and no

[2]If the critical level is negative, this case is moot.

information – to the very common situation where some customers are informed and the rest are not.

To answer those questions, we study a single-server queue as in Hassin (1986) but with a middle ground by assuming a mix of two streams of customer arrivals who are different in their information structures. Specifically, the server posts information on the actual queue length in real time. One stream of customers, which we call *informed*, makes the decision to join or balk on the basis of the real-time information about delays. The other stream, which we call *uninformed*, do not have access to this real-time information and base their join-or-balk decision on their past experiences of congestion levels. We assume the fraction of informed customers, which we also refer to as *information level*, is *exogenous*, and we study comparative statics with respect to the information level, namely the influence of an larger informed fraction (i.e., growing information prevalence) on system performances such as throughput and social welfare.

Given the difference in the delay information they possess, the two streams of customers have totally different self-interested joining behavior. Informed customers use a threshold policy: if the queue is observed to be shorter than a particular threshold, they join it; otherwise, they balk. Uninformed customers, on the contrary, are only aware of the expected waiting time through their past experiences and randomize their decisions between joining and balking. Although informed customers use the same state-dependent threshold strategy as if they were in an observable queue (cf. Naor 1969), the presence of uninformed customers undoubtedly influences the probability that an informed customer will join the queue. On the one hand, this interaction is not captured by either observable or unobservable models. On the other hand, our results reveal that service providers who ignore this interaction between the two segments may miss the opportunity to achieve better system performance measures. In particular, we show that unless the customer volume is extremely low, it is possible to improve either throughput or social welfare when only a fraction of customers are informed. Moreover, these system performance measures depend crucially on the equilibrium joining behavior of *uninformed* customers.

We show that the ubiquity of delay information may have a positive or negative impact on the system throughput. In particular, we prove that there are two critical levels of offered loads. If the offered load is above (or below) the higher (or lower) one, the throughput always increases (or decreases) in the information level. These results are consistent with the comparison of the full- and no-information models in Chen and Frank (2004). However, we also show that if the offered load falls in the *intermediate* range between the two critical levels, the throughput is always *unimodal* in the information level. In addition, the throughput reaches its maximum at the information level in which

all uninformed customers are about to adopt an always-join strategy. This finding implies that treating all customers equally either as informed or as uninformed may fail to realize the potential value of effective information control.

Notice that system throughput would steadily accumulate at the service rate as long as the server is not idle. Thus, maximizing throughput is equivalent to minimizing system idleness. Idleness stems from either inadequate service requests (namely, the arrival rate is low), which we call the mean effect, or an inter-temporal mismatch between capacity and demand, which we call the variability effect. Although information prevalence can counteract these two effects in some situations, it may also fail to do so. If the arrival rate is large relative to the service rate, the mean effect is of little concern and the variability effect can be mitigated as more customers become informed and join the line immediately when it is short. In contrast, if the offered load is not high, the server is more likely to experience too few service requests. Thus, it is essential to improve the average joining probability of each customer. Although information prevalence marginally decreases each informed customer's probability of joining, it dramatically motivates the uninformed ones under a relatively low customer load, due to a shorter expected delay. As a result, the net effect is that the average joining probability of the entire customer pool increases in the information level. However, if all uninformed customers have already adopted an always-join strategy, they cannot be further stimulated. Consequently, system throughput starts to suffer from a further increase in the information level, which in fact turns away customers who would join the line if they could not see its length. In particular, if the offered load is so low that all customers are able to receive positive utility even without real-time information, any marginal increase in the information level only hurts the throughput.

Contrary to the conventional wisdom that congestion information always improves social welfare, we further demonstrate that social welfare is *unimodal* in the information level when the system experiences a sufficiently high offered load (only when the offered load is relatively low does information prevalence always benefit social welfare). This is because growing information prevalence has both positive and negative impacts on social welfare. On the positive side, if the system congestion is visible to customers, system capacity can be more efficiently matched with service requests inter-temporally because potential informed customers seek service only when the queue is short enough to yield positive utilities. However, informed customers' self-interested joining behavior may overload the system, especially when the service requests are overwhelming. The overall effect on social welfare depends on the interactions between the informed and uninformed customer segments. Under a large offered load, uninformed customers expect

a long line and join the queue with a very low probability. This disincentive helps mitigate the congestion. Furthermore, as the fraction, hence the number, of informed customers increases, those desirable positions with only a few people waiting ahead are more likely to be taken instantaneously by informed customers. Therefore, the queue has a tendency to grow longer in the information level. In response, the remaining uninformed customers would be even less likely to join the line. The declining incentive for uninformed customers who earn *zero* utilities in turn frees up the capacity to serve more informed customers, who contribute *positive* utilities to welfare. Nonetheless, when a large proportion of a high volume of customers are informed, uninformed individuals eventually lose interest in the service and all choose to balk. Without uninformed customers' concession, the system suffers from escalating externality inflicted by the increasing number of informed customers and hence social welfare deteriorates.

Our results highlight the fact that some degree of information heterogeneity in real-time delay information in the population can lead to more *efficient* outcomes than information homogeneity in terms of system throughput or social welfare. The presence of uninformed customers or their behavior does not necessarily harm the system. In fact, it improves the system throughput when the system experiences low offered loads and increases social welfare when the system experiences high offered loads.

To evaluate the effect of information heterogeneity, we study comparative statics by assuming an exogenous fraction of uninformed customers. However, we caution that all customers will choose to be informed if they are fully rational and there is no cost of being informed. Assuming all customers are rational, we thus discuss how to achieve a desirable degree of information heterogeneity by charging an information access fee. Our results from the base model imply the potential value of intentionally concealing delay information from certain customers. For instance, call centers may consider making delay announcements only to premium customers. In that situation, it is also plausible that informed customers such as premium ones, and uninformed customers such as regular ones, may value the service differently and incur different unit delay costs. We thus examine an extension with heterogeneous customer characteristics in service reward and delay cost. We find that system throughput and welfare can still be unimodal in the information level.

Literature Review

Literature in the influence of delay information on customer behavior dates back to Naor (1969). The author argues that in an observable service system, customer self-

interested joining decisions, which ignore the negative externality on later arrivals, overload the system and result in a deviation from social optimality. Hassin and Haviv (2003) comprehensively summarize various extensions to Naor (1969). Hassin (1986) studies a revenue-maximizing server who has the option of completely suppressing the information on real-time queue length. The author shows that when a revenue maximizer prefers to reveal the queue length, so does a social planner.

Customers may sometimes not be able to observe system states directly but have to rely on delayed information announced by service providers. One example is delay announcements in call centers. Whitt (1999) argues that informing customers about anticipated delays can effectively reduce customer abandonment. Guo and Zipkin (2007) study the effects of delaying information with different degrees of precision: no information, the queue length, and the exact waiting time. They find that exact delay information may either improve or hurt social welfare because all customers are not equally patient. This finding is further strengthened by Guo and Zipkin (2009). These papers on delay announcement all implicitly assume that service providers offer truthful information. However, customers are often unable to verify the announced congestion information. Allon et al. (2011) model customer strategic response to provider's unverifiable delay information and characterize equilibrium signaling languages that emerge between the service provider and her customers. Allon and Bassamboo (2011) further reveal that delaying the announcements about waiting times can make the announced information more credible.

There is an emerging stream of literature on behavioral queues. Plambeck and Wang (2013) show how customers' lack of self-control and naivete affect optimal pricing and scheduling in a service system. Huang et al. (2013) study canonical service models with boundedly rational customers. They find that for observable queues with endogenized pricing, bounded rationality results in a loss of revenue and welfare. Cui and Veeraraghavan (2014) study a queue that serves a pool of customers who may have arbitrarily misinformed beliefs about the service rate. The authors show that revealing the service information to consumers can benefit revenues but may hurt individual welfare or social welfare. Another stream of research in this behavioral queue literature studies herding in queues. Veeraraghavan and Debo (2009, 2011) and Debo and Veeraraghavan (2014) study customer inferences about service quality through observation of the length of waiting lines, which may lead to herding in queues.

All the papers referred to above assume that customer perceptions of delay information are homogeneous: i.e., either no one has access to the information or all receive the same types of information. However, as we argued before, it may be unrealistic to

assume that all customers are aware of system congestion even though such information is available through many channels. In contrast to previous work, we consider customer heterogeneous perceptions of delay information; i.e., only a fraction of customers can obtain information about the real-time queue length. Our work focuses on the interaction between informed and uninformed customers and the resulting system performance measures.

Most relevant to our work is Hassin and Roet-Green (2013), which models rational customer decisions among three actions: join, balk, or incur a *hassle cost* to inspect the queue length before making a join-or-balk decision. The authors prove the existence of a customer equilibrium strategy, which is effectively a randomization of the three possible actions. The authors show that the service provider can have a higher throughput if customers must incur a hassle cost to inspecting the queue so that only a fraction of customers are informed of the queue length. They show that social welfare is maximized when the inspection is costless and thus all customers are informed. In contrast, we show that our base model can be adapted to account for the situation in which the service provider charges an *information fee* (considered as a payment transfer between the provider and customers) and the customers are completely rational in deciding between paying an information fee to be informed or not paying and staying uninformed. Therefore, results from our base model imply that charging an information fee can improve system throughput or social welfare. Our extension to rationalizing customers' uninformed behavior with an information fee may provide an alternative way to prove the equilibrium existence result in Hassin and Roet-Green (2013). In addition, for our setting, we identify the behavior of uninformed customers as the driving force of various results and provide explanations.

3.2 Model Setup

A single-server facility expects a stream of customers who arrive one at a time according to a Poisson process with rate Λ. Customers are risk-neutral and are served on a first-come-first-served basis. The service time of each customer is independent and exponentially distributed with mean $1/\mu$. We denote $\rho \equiv \Lambda/\mu$ as the offered load of the system. We assume that the admission fee to the facility is an irrelevant factor in a customer's join-or-balk decision and is thus scaled to zero; as a result, social welfare does not involve the service provider's profit. Such services may include boarder crossings, driving on

highways, and rides at Disney World[3]. Upon completion of the service, a customer receives a reward R. During the sojourn time in the system, a linear waiting cost with marginal rate c is incurred. We further assume that the service reward is sufficient to offset the waiting cost when there is no line upon arrival, i.e., $R\mu \geq c$. The same assumption is commonly made in the literature (e.g., Naor 1969, Hassin 1986). If a customer chooses to balk, she receives zero utility. Moreover, we assume that customers do not renege.

There are two streams of customers. The *informed* stream checks the real-time information on the queue length, $Q(t)$, and takes that into account when making an individual join-or-balk decision. The other *uninformed* stream ignores or is not able to obtain the real-time information. Thus the uninformed customers have to rely on their expectations of queue lengths, based on their past experiences with the congestion levels, in deciding whether to join or balk. The fraction of informed customers in the whole population is denoted by an exogenous parameter $\gamma \in [0, 1]$, which measures the real-time information level in the society. We denote by λ_{I} and λ_{U} the arrival rates of informed and uninformed customer streams respectively. Then, we have

$$\lambda_{\text{I}} \equiv \gamma\Lambda \quad \text{and} \quad \lambda_{\text{U}} \equiv (1-\gamma)\Lambda.$$

It is natural to assume that customers in each stream have their private knowledge of their own service reward R and unit delay cost c, which are homogeneous for both streams in the base model. For equilibrium analysis, we assume that all system parameters are known to the uninformed customers.[4] It may be restrictive to assume that uninformed customers know the fraction of informed customers γ. In Section 3.5, we relax this particular informational requirement by endogenizing the information level as an outcome of rational customers' utility maximization given an information fee. Some recent papers

[3]Although a customer needs to buy a day pass, there is no additional charge for any ride or attraction. Therefore, price is not likely to be a factor when a customer chooses a ride.

[4]To reach a system equilibrium, the minimum information assumption for uninformed customers may be that they only know their own service reward R and delay sensitivity c, but they react actively to their past experiences with the system. The segment of uninformed customers, as a whole, would increase (decrease) their joining probability if $cE(\overline{W}) < (>)R$, where $E(\overline{W})$ is the expected delay from their past experiences over a short term. Because the expected wait time is increasing in the joining probability of uninformed customers (see Lemma 3.1), through a dynamic process of interacting with the system, uninformed customers would reach a unique equilibrium over time where they have no incentive to deviate from their choice of joining probability. No doubt this minimum information structure requires customers' repeated interactions with the system, which may be reasonable for settings such as frequent highway commuters who respond to past traffic conditions. In other settings, to reach an equilibrium, we do require the uninformed to know all system parameters, i.e., the arrival rate Λ, service rate μ, the fraction of the informed customers γ, and the informed customers' service reward and delay sensitivity, beyond their private knowledge of their own.

also consider relaxation of those restrictive informational requirements; see, e.g., Debo and Veeraraghavan (2014); Cui and Veeraraghavan (2014).

3.3 Join-or-Balk Decisions

We next discuss how the two customer streams make their joining decisions in equilibrium. Since the two streams possess different levels of information about queue length, they have different strategies.

3.3.1 State-Dependent Decisions of Informed Customers

Informed customers have private knowledge of their own service reward and sensitivity to delays, and they make contingent decisions according to the system state. Upon arrival, knowing that i customers are in the system, including the one receiving service if any, an informed customer joins the queue if and only if the expected net value $R-c(i+1)/\mu \geq 0$, i.e., $i+1 \leq R\mu/c$. For notation simplicity, $\lfloor x \rfloor$ denotes the largest integer that is less than or equal to x, and $\langle x \rangle \equiv x - \lfloor x \rfloor \in [0,1)$ denotes the *fractional part* of real number x. Therefore, there is a threshold

$$n \equiv \lfloor \nu \rfloor, \text{ where } \nu \equiv \frac{R\mu}{c} \geq 1,$$

such that an informed customer arriving at time t joins the queue if and only if she observes $Q(t) < n$, and otherwise balks. In other words, $n-1$ is the maximum queue length beyond which joining the queue would lead to negative utility for an informed customer. Our model of informed customer behavior is in the same vein as those in observable queues (e.g., Naor 1969, Hassin 1986).

3.3.2 Equilibrium Mixed Strategies of Uninformed Customers

The uninformed customers are unaware of the real-time queue length. Specifically, uninformed customers' strategy can be described by a fraction $q \in [0,1]$. We can interpret q as either the proportion of all uninformed customers who seek service or the probability that each uninformed customer joins the queue. Let $p_i(q)$ denote the probability that there are i customers waiting in the queue in the steady state when the joining probability of uninformed customers is q. The *collective* behavior of the informed and uninformed

streams jointly determines the following balance equations of the process:

$$(\lambda_{\text{I}} + q\lambda_{\text{U}})p_i(q) = \mu \cdot p_{i+1}(q) \quad \text{if } 0 \le i < n, \tag{3.1}$$

$$q\lambda_{\text{U}} \cdot p_i(q) = \mu \cdot p_{i+1}(q) \quad \text{if } i \ge n. \tag{3.2}$$

Thus, the probability density function of the queue length can be written as:

$$p_i(q) = \begin{cases} (\rho_{\text{C}}(q))^i \, p_0(q) & \text{if } 0 \le i < n, \\ (\rho_{\text{C}}(q))^n \, (\rho_{\text{U}}(q))^{i-n} \, p_0(q) & \text{if } i \ge n, \end{cases} \tag{3.3}$$

where, for convenience of notation,

$$\rho_{\text{C}}(q) \equiv \frac{\lambda_{\text{I}} + q\lambda_{\text{U}}}{\mu} \text{ and } \rho_{\text{U}}(q) \equiv \frac{q\lambda_{\text{U}}}{\mu}. \tag{3.4}$$

(We may suppress the dependence of $\rho_{\text{C}}(q)$ and $\rho_{\text{U}}(q)$ on q to further simplify the notation.) Using $\sum_{i=0}^{\infty} p_i(q) = 1$, we can derive the idle probability

$$p_0(q) = \left(\sum_{i=0}^{n-1} (\rho_{\text{C}})^i + \frac{(\rho_{\text{C}})^n}{1-\rho_{\text{U}}} \right)^{-1} = \left(\frac{1-(\rho_{\text{C}})^n}{1-\rho_{\text{C}}} + \frac{(\rho_{\text{C}})^n}{1-\rho_{\text{U}}} \right)^{-1}. \tag{3.5}$$

Therefore, the expected sojourn time W in the steady state is

$$W(q) = \sum_{i=0}^{\infty} \frac{i+1}{\mu} p_i(q) = \frac{p_0(q)}{\mu} \left[\frac{1-(\rho_{\text{C}})^n}{1-\rho_{\text{C}}} + \frac{\rho_{\text{C}}}{(1-\rho_{\text{C}})^2} + (\rho_{\text{C}})^n \left(\frac{1-n}{1-\rho_{\text{C}}} - \frac{1}{(1-\rho_{\text{C}})^2} + \frac{1}{(1-\rho_{\text{U}})^2} + \frac{n}{1-\rho_{\text{U}}} \right) \right]. \tag{3.6}$$

Then, uninformed customers with a joining strategy q receive utility $R - cW(q)$ on average. If there were only uninformed customers, it is obvious that ceteris paribus, the expected sojourn time would increase as q becomes larger. However, we have two customer streams. As uninformed customers join the line more frequently, informed customers are less likely to join, and that alleviates the congestion. The next lemma confirms that the former effect dominates the latter.

Lemma 3.1 *For any given $\gamma \in [0, 1)$, the queue length $Q(q)$ in the steady state is stochastically increasing in q and thus the expected sojourn time $W(q)$ is strictly increasing in q.*

Following Hassin and Haviv (2003), we can determine the unique equilibrium joining probability of uninformed customers by the strict monotonicity of $W(q)$. Specifically, let $q^* \in [0, 1]$ be the equilibrium joining probability of the uninformed customers. If

$R - cW(0) \leq 0$, an uninformed customer earns a non-positive utility even if no other uninformed ones join. Therefore, $q^* = 0$. If $R - cW(1) \geq 0$, uninformed customers receive non-negative net benefits even if they all join. Therefore, $q^* = 1$. Otherwise, a mixed strategy will be used. Assume that all uninformed customers join the queue with a probability $0 < q < 1$ and $R - cW(q) > 0$. Since each customer is an infinitesimal entity, an uninformed customer can unilaterally improve her own utility by joining more frequently. Therefore, it must be that $R - cW(q^*) = 0$ in equilibrium. The next proposition summarizes the equilibrium joining strategy by uninformed customers. To draw a parallel between the joining strategies of informed and uninformed customers, we present the results in terms of the expected queue length $\mathbb{E}[Q(q)]$.

Proposition 3.2 *Fix $\gamma \in [0, 1)$. There exists a unique equilibrium joining strategy q^* for uninformed customers.*

(i) (Always Balk: No Participation) $q^* = 0$ *if and only if* $\mathbb{E}[Q(0)] \geq \nu$, *i.e.,* $cW(0) \geq R$.

(ii) (Always Join: Full Participation) $q^* = 1$ *if and only if* $\mathbb{E}[Q(1)] \leq \nu$, *i.e.,* $cW(1) \leq R$.

(iii) (Randomize between Balking and Joining: Partial Participation) $q^* \in (0, 1)$ *must satisfy* $\mathbb{E}[Q(q^*)] = \nu$, *i.e.,* $cW(q^*) = R$, *if and only if* $\mathbb{E}[Q(0)] < \nu < \mathbb{E}[Q(1)]$, *i.e.,* $cW(0) < R < cW(1)$.

From the above proposition, we can further analytically identify the system primitives, under which the uninformed customers always balk or join in equilibrium, by exploring the condition of $cW(0) \geq R$ or $cW(1) \leq R$ respectively. In the rest of the primitive space, uninformed customers randomize their decisions between joining and balking in equilibrium.

Corollary 3.3 (NO PARTICIPATION) *All uninformed customers always balk at the queue in equilibrium, i.e., $q^* = 0$, if and only if $1 \geq \gamma \geq \gamma_0^*(\rho, \nu) \equiv \frac{y^*(\nu)}{\rho}$, where $y^*(\nu) \geq 0$ is the unique solution to $n + 1 + \frac{1}{1-y} - \frac{n+1}{1-y^{n+1}} = \nu$.*[5]

Corollary 3.4 (FULL PARTICIPATION) *All uninformed customers always join the queue in equilibrium; i.e., $q^* = 1$, if and only if*

$$1 \geq \gamma \geq \gamma_1^*(\rho, \nu) \equiv 1 - \frac{1}{\rho} + \frac{2}{\rho}\left(\langle\nu\rangle + \sqrt{\langle\nu\rangle^2 + 4L(\rho, \nu)}\right)^{-1},$$

[5]The cut-off point $\gamma_0^*(\rho, \nu)$ can be shown to be always non-negative but can be larger or smaller than 1. If $\gamma_0^*(\rho, \nu) > 1$, the case $q^* = 0$ is moot; i.e., there exists no $\gamma \in [0, 1]$ such that $q^* = 0$.

where $L(\rho,\nu) \equiv \frac{\langle\nu\rangle(\rho-1)\rho^n+\nu-\nu\rho+\rho^n-1}{(1-\rho)^2\rho^n} = \frac{\nu-\langle\nu\rangle\rho^n-\sum_{i=0}^{n-1}\rho^i}{(1-\rho)\rho^n} \geq 0$ *and* $\langle v\rangle \equiv v-n$.[6]

Corollaries 3.3 and 3.4 show that uninformed customers would always join or balk when the information level is beyond a certain threshold. Nevertheless, depending on the system offered load, only one threshold, $\gamma_0^*(\rho,\nu)$ or $\gamma_1^*(\rho,\nu)$, can exist in the range of $[0,1]$ for a given set of system primitives. The next proposition describes how the offered load ρ and information level γ jointly determine the equilibrium joining behavior of uninformed customers.

Theorem 3.5 (EQUILIBRIUM STRATEGY OF UNINFORMED CUSTOMERS) *For given* ρ *and* ν, *define* $\underline{\rho} \equiv 1-1/\nu$ *and* $\overline{\rho} \equiv y^*(\nu)$ *respectively. The equilibrium joining probability* q^* *of uninformed customers depends on* ρ *and* γ *in the following way:*

(i) (Always Full Participation) *If* $0 \leq \rho < \underline{\rho}$, $q^* = 1$ *for all* $0 \leq \gamma \leq 1$.

(ii) (Partial to Full Participation) *If* $\underline{\rho} \leq \rho \leq \overline{\rho}$, $q^* \neq 0$ *for all* $0 \leq \gamma \leq 1$. *In particular,* $0 < q^* < 1$ *for* $0 \leq \gamma < \gamma_1^*(\rho,\nu)$ *and* $q^* = 1$ *for* $\gamma_1^*(\rho,\nu) \leq \gamma \leq 1$, *where* $\gamma_1^*(\rho,\nu) \in [0,1]$.

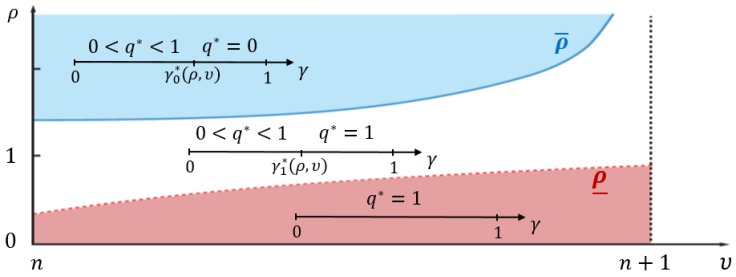
(iii) (Partial to No Participation) *If* $\rho > \overline{\rho}$, $q^* \neq 1$ *for all* $0 \leq \gamma \leq 1$. *In particular,* $0 < q^* < 1$ *for* $0 \leq \gamma < \gamma_0^*(\rho,\nu)$ *and* $q^* = 0$ *for* $\gamma_0^*(\rho,\nu) \leq \gamma \leq 1$, *where* $\gamma_0^*(\rho,\nu) \in [0,1]$.

We next use Figure 3.1 to illustrate the results in Proposition 3.5. The vertical axis represents the system offered load $\rho = \Lambda/\mu$, and the horizontal axis is $\nu = R\mu/c$, which ranges from n inclusive to $n+1$ exclusive. The lower dashed and upper solid curves correspond to the two thresholds $\underline{\rho}$ and $\overline{\rho}$ respectively. Moreover, it can be shown that the solid curve always stays above the dashed one and $\overline{\rho}$ approaches infinity when ν tends to $n+1$. The two offered load thresholds $\underline{\rho}(\nu)$ and $\overline{\rho}(\nu)$ divide uninformed customer equilibrium strategies into three types in the primitive space of (ρ,ν). We discuss each as follows.

In the area below $\underline{\rho}$, the offered load is extremely low and the expected sojourn time for uninformed customers is very short. Even if no one observes the queue length, namely $\gamma = 0$, all uninformed customers choose to join the queue, i.e., $q^* = 1$. As the information level γ increases by an infinitesimal amount, a small proportion of customers change from uninformed to informed. These converted customers now join the queue only if they observe that $Q(t) < n$, rather than definitely as before. The system congestion is

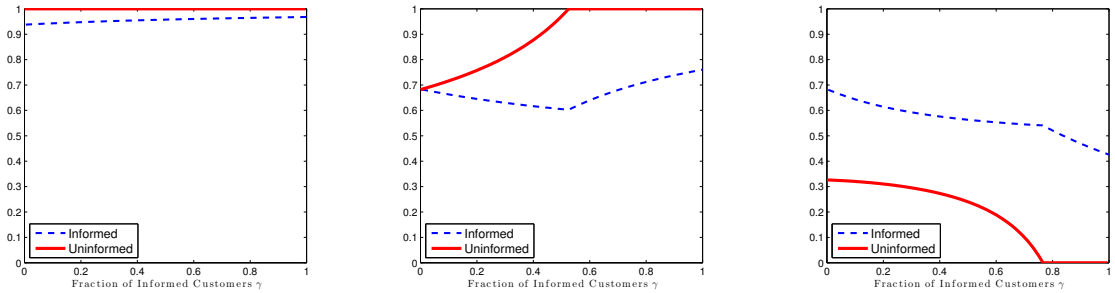
[6]The cut-off point $\gamma_1^*(\rho,\nu)$ can be negative. If $\gamma_1^*(\rho,\nu) < 0$, then the case $q^* = 1$ holds for all $\gamma \in [0,1]$. Moreover, $\gamma_1^*(\rho,\nu)$ can be larger than 1. If $\gamma_1^*(\rho,\nu) > 1$, the case $q^* = 1$ is moot, i.e., there exists no $\gamma \in [0,1]$ such that $q^* = 1$.

Figure 3.1: **Illustration for Equilibrium Behavior of Uninformed Customers**



hence slightly alleviated. The reduced delay, on the one hand, reinforces the remaining uninformed customers' incentives to join, with the result that they still all join the queue as the information level γ grows. This also intuitively justifies the validity of Corollary 3.4. On the other hand, the slightly reduced congestion also increases the probability that informed customers will see a short queue. Therefore, the chance of an informed customer joining the line slowly increases in γ when $q^* = 1$. Figure 3.2(a) depicts q^* as a function of γ in such a scenario when $\rho = \Lambda/\mu = 0.5$, the service reward $R = 4$, and marginal waiting cost $c = 1$. In this case, uninformed customers always join the queue in equilibrium regardless of the information level as shown in Theorem 3.5(i) and informed customers' probability of joining rises slightly.

Figure 3.2: **Joining Probabilities of Informed and Uninformed Customers** ($\mu = 1, R = 4,$ **and** $c = 1$)



(a) $\rho = \Lambda/\mu = 0.5$ (b) $\rho = \Lambda/\mu = 1.1$ (c) $\rho = \Lambda/\mu = 2.3$

As the offered load gradually increases to intermediate levels between the dashed and solid lines in Figure 3.1, a large fraction or all of the uninformed customers join the queue in equilibrium depending on the information level γ (see Theorem 3.5(ii)). In this case, the total customer volume is relatively modest relative to the service speed. Even if a considerable number of customers are informed, e.g., $\gamma \to 1^-$, the expected sojourn time for uninformed customers is still endurable, since the absolute number of informed customers is not very large and they enter the queue only when it is short. Therefore,

uninformed customers line up regardless. For instance, in Figure 3.2(b), $q^* = 1$ if γ is close to 1 when the offered load ρ equals 1.1. However, as γ decreases, e.g., $\gamma \to 0^+$, the number of uninformed customers rises. Every uninformed customer who is risk-neutral has to seek service less often to cope with the increasing negative externalities from her uninformed peers who join blindly at a certain probability. This thus reduces uninformed customers' incentive to join the line. Moreover, because uninformed customers join less as γ declines, the queue is more likely to be shorter than n, and that increases the probability that informed customers will join. As the information level γ approaches 0, uninformed customers are discouraged from lining up but informed customers are slightly encouraged, compared to when $q^* = 1$. Visually, these facts correspond to the increasing solid curve and declining dashed curve respectively up to $\gamma = 0.51$ in Figure 3.2(b). Beyond $\gamma = 0.51$, the pattern is similar to Figure 3.2(a).

While the offered load rises to excessive levels, i.e., above the solid curve in Figure 3.1, only a small fraction or none of the uninformed customers join the queue in equilibrium depending on the information level γ (see Theorem 3.5(iii)). In this case, the system confronts enormous customer volumes. With no knowledge of the real queue length, uninformed customers expect a long line and have very little incentive to join. With such a large customer volume, the advantage of real-time congestion is clear: any available spot with fewer than n people ahead will be taken quickly. As the number of informed customers increases, this effect become more salient and the queue becomes even longer. Informed customers observe a short queue less frequently and have to balk more often as γ increases to 1. For uninformed customers, the incentive to join the line diminishes in γ until it vanishes completely. For the same reason, a further increment in the number of informed customers beyond the vanishing point where q^* hits 0 only causes those appealing spots to be occupied even faster and the queue to be even longer. Hence, q^* remain at 0 after the information level exceeds $\gamma_0^*(\rho, \nu)$ as shown in Corollary 3.3. Figure 3.2(c) displays such dynamics when the offered load $\rho = 2.3$.

3.4 Impacts of Heterogeneous Information

To answer the questions raised in the introduction, we investigate how a marginal increase in the information level would affect the system performance measures. Specifically, we are interested in the system throughput and social welfare in equilibrium. The system throughput, denoted by λ, includes effective arrival rates of both informed and uninformed streams. That is,

$$\lambda(q) = \left(\sum_{i=0}^{n-1} p_i(q) \right) \lambda_{\text{I}} + q\lambda_{\text{U}}. \tag{3.7}$$

Likewise, social welfare is composed of contributions from both segments. Let $S_{\mathrm{I}}(q)$ and $S_{\mathrm{U}}(q)$ be the total net utilities of informed and uninformed customers respectively. We have

$$S_{\mathrm{I}}(q) = \left[\sum_{i=0}^{n-1} p_i(q)\left(R - c\frac{i+1}{\mu}\right)\right] \cdot \gamma\Lambda$$

and

$$S_{\mathrm{U}}(q) = \left[q\sum_{i=0}^{\infty} p_i(q)\left(R - c\frac{i+1}{\mu}\right)\right] \cdot (1-\gamma)\Lambda = q(R - cW(q)) \cdot (1-\gamma)\Lambda.$$

Social welfare is thus the sum of all customer net utilities, namely,

$$S(q) = S_{\mathrm{I}}(q) + S_{\mathrm{U}}(q). \tag{3.8}$$

Although informed customers use the same state-dependent threshold strategy as if they were in an observable queue (cf. Naor 1969), their contributions to the effective arrival rate and social welfare also hinge on uninformed customer behaviors through $p_i(q)$, $i = 0, \ldots, n-1$. On the one hand, this interaction, which is not captured by either the observable model or the unobservable one, implies the decisive influence of uninformed customer behavior in a general circumstance. On the other hand, it also increases the technical difficulty of the analysis. One may be tempted to seek for a general closed-form expression for the equilibrium joining probability q^* of uninformed customers as a function of γ and conduct comparative statics of $\lambda(q^*(\gamma))$ and $S(q^*(\gamma))$ on the information level γ. However, it is a daunting, if not impossible, task. That is because the equilibrium joining probability $q^* \in (0,1)$ is characterized by a high-order polynomial function: $cW(q) = R$, where $W(q)$ is specified in Eq.(3.6). According to the Abel-Ruffini Theorem, this type of equation in general has no algebraic solution in radicals. Although the insolvability of q^* hinders a direct approach to analyzing the system, we will show monotonicity properties of equilibrium throughput and social welfare with respect to the information level γ by taking an indirect approach. Somewhat surprisingly, their monotonicity properties are uniquely determined by the *type* of equilibrium joining strategy that uninformed customers adopt, namely, always-balk (i.e., $q^* = 0$), always-join (i.e., $q^* = 1$), or randomization between balking and joining (i.e., $0 < q^* < 1$), which is an outcome of interactions between informed and uninformed customers.

3.4.1 Throughput

Chen and Frank (2004) compare the throughput in the observable (i.e., $\gamma = 1$) and unobservable (i.e., $\gamma = 0$) queues. They demonstrate that there is a unique critical level ρ^* such that if $\rho > \rho^*$, the throughput of the observable queue is more than that of the

unobservable queue, and the converse holds if $\rho < \rho^*$. Their result provides an answer to the comparison of $\lambda\left(q^*(\gamma=0)\right)$ and $\lambda\left(q^*(\gamma=1)\right)$, but does not reveal the marginal effect of information on the equilibrium throughput $\lambda(q^*(\gamma))$ in general for $\gamma \in [0,1]$. We will characterize the monotonicity of the system equilibrium throughput λ in the information level γ in this subsection.

We notice that it is difficult to directly analyze the throughput formula $\lambda(q)$ in (3.7). However, if we take a summation of all system states in (3.1) and (3.2), we observe that

$$\sum_{i=1}^{n-1}(\lambda_{\text{I}}+q\lambda_{\text{U}})p_i(q)+\sum_{i=n}^{\infty}q\lambda_{\text{U}}p_i(q)=\mu\sum_{i=0}^{\infty}p_{i+1}(q)\Longleftrightarrow\lambda(q)=\left(\sum_{i=0}^{n-1}p_i(q)\right)\lambda_{\text{I}}+q\lambda_{\text{U}}=\mu(1-p_0(q)).$$

In other words, the system throughput equals the service rate minus the vacant capacity due to idleness. Therefore, we can explore the monotonicity of the equilibrium throughput via the probability of idleness in equilibrium.

Lemma 3.6 *If $q^* \in [0,1)$, the probability $p_0(q^*(\gamma))$ that the server is idle strictly decreases in γ.*

The above result means that as long as uninformed customers do not take the full-participation strategy in equilibrium, the server is less likely to be idle as the real-time congestion information becomes more prevalent. The declining idleness of the server implies the growth in throughput as summarized in the next theorem.

Theorem 3.7 (Comparative Statics of Throughput) *(i) If $0 \le q^* < 1$, the throughput $\lambda(q^*)$ is strictly increasing in γ.*

(ii) If $q^* = 1$, the throughput $\lambda(q^*)$ is strictly decreasing in γ.

Theorem 3.7 says that system throughput benefits marginally from growing information prevalence unless all uninformed customers choose to join the queue in equilibrium. However, the negative effect of information on throughput can happen in a large range. We use an example to illustrate this.

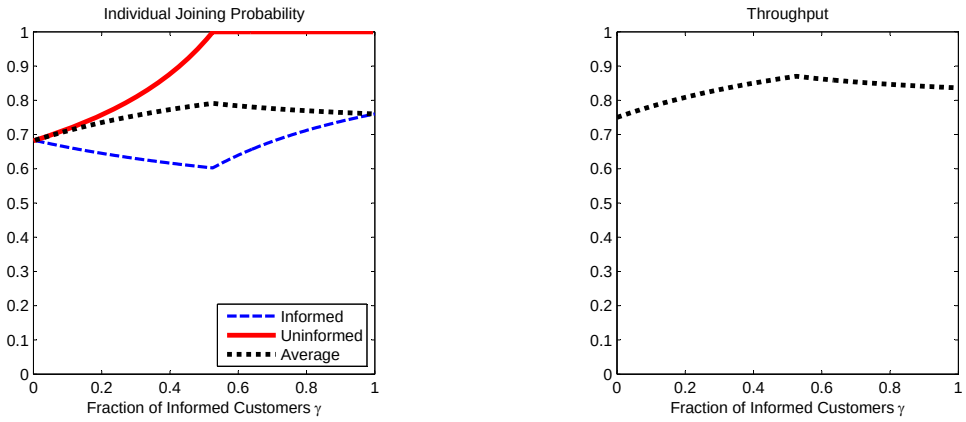
Example 3.1 *Consider a system with the offered load ρ close to 1. From Corollary 3.4, $\lim_{\rho\to 1} L(\rho,\nu) = \langle\nu\rangle n + n(n-1)/2$ by L'Hôpital's rule and hence, $\lim_{\rho\to 1}\gamma_1^*(\rho,\nu) \le 2/\sqrt{2n(n-1)}$, which can be further bounded by $1/\sqrt{3} \approx 57.7\%$ when $n-1 \ge 2$.*

Corollary 3.4 confirms that if $\gamma \ge \gamma_1^*(\rho,\nu)$, uninformed customers always choose to join the queue. Theorem 3.7 implies that if informed customers have a joining threshold no less than 2, throughput will always suffer if there are more than 58% of informed customers in the population. This simple example shows clearly that the system throughput

can easily suffer from real-time congestion information if too many customers are informed. This phenomenon is a result of equilibrium responses by uninformed customers to the heterogeneous availability of information. Hence, it is not covered by the conventional observable and unobservable frameworks. ■

We next explain the intuition of Theorem 3.7 through its connection to Lemma 3.6. As we have established, maximizing the throughput is equivalent to minimizing the server's probability of being idle, namely, $p_0(q^*)$. In principle, there are two reasons for idleness: inadequate service requests (viz. low arrival rates) and an intertemporal mismatch between capacity and demand, i.e., the mean effect and the variability effect. Inadequate service requests result from long times between arrivals relative to the mean service time. In this case, the server often completes tasks at a rate higher than that of arrivals. Therefore, the server is exposed to a high risk of idleness. The mismatch between capacity and demand stems from the uncertainty about service times and arrivals. Service requests may not always arrive at the moment the server is idle or the queue is short. As we shall discuss, which effect is more dominant in causing idleness depends on the offered load.

Figure 3.3: Example: $\mu = 1, R = 4, c = 1$, and $\Lambda = 1.1$



(a) Joining Probability

(b) Throughput

When the offered load is relatively low, the customer arrival rate is low, and thus the server is very likely to complete all tasks before another service request arrives. Hence, too few service requests is the first-order effect that gives rise to idleness. In order to improve throughput, the provider has to increase the average probability that an individual customer will join the line. This rationale can also be seen mathematically.

Rearranging (3.7), we write the throughput as

$$\lambda(q^*(\gamma)) = \left(\sum_{i=0}^{n-1} p_i(q^*(\gamma))\gamma + q^*(\gamma) \cdot (1-\gamma) \right) \Lambda,$$

where the term in the brackets represents the average joining probability of the entire population. We next examine the role of information in incentivizing customer entrance of the queue. Note that when the offered load is low, the queue is expected to be short regardless of the information level γ. Although a marginal change in the information level does affect the probability that the queue will be shorter than n and thus the chance that an informed individual will join the queue, such an effect is very marginal due to the low offered load. Therefore, the incentive for informed customers is not very sensitive to information level changes, either upward or downward. In contrast, uninformed customers are more sensitive to the change in information prevalence. As more uninformed individuals become informed, the number of uninformed customers decreases. Moreover, those who used to be uninformed but are now informed also join less often because they anticipate positive, instead of zero, net utility to justify their participation. These two effects together provide the remaining uninformed customers with a stronger incentive to join. Consequently, as information becomes more ubiquitous, uninformed customers are much more motivated to join the queue. Note that their enthusiasm reduces the chance that the informed customers will join. But as we argued before, this effect is marginal due to the low offered load. Combining both segments, we can see that the average customer joining probability rises as a function of the information level γ, with the uninformed segment being more significantly incentivized and the informed segment being slightly discouraged. The dotted curve in Figure 3.3(a) shows the average joining probability of the entire customer pool. As we see, it increases up to $\gamma = 0.51$, at which point uninformed customers start to join definitely. Unfortunately, the remaining uninformed customers cannot be further stimulated if all of them have chosen to join. Therefore, when q^* reaches 1, the provider loses its beneficial leverage in motivating uninformed customers through information. As more uninformed customers become informed after q^* hits 1, they join less than when they were uninformed. Consequently, the average joining probability goes down, as shown by the dotted curve in Figure 3.3(a) beyond $\gamma = 0.51$. Figure 3.3(b) displays the change in throughput as a function of γ. The pattern in Figure 3.3(b) exactly matches that of the average joining probability in Figure 3.3(a).

When the offered load is high enough, the customer arrival rate is high, and thus there are enough service requests. Yet, an intertemporal mismatch between capacity

and arrivals due to system variability becomes the primary reason for server idleness. In this situation, increasing the availability of real-time information is an effective strategy. That is because the more informed customers there are, the more quickly the desirable positions with fewer than n customers ahead are taken. This reduces the likelihood that the server being idle, thus improving the throughput.

In summary, increasing the availability of information improves system throughput unless uninformed customers all join the queue. However, the reason for the phenomenon may be different for various offered loads. Under a low offered load, growing information prevalence slightly reduces each informed customer's probability of joining the queue, but it dramatically motivates risk-neutral uninformed customers to join. The average joining probability thus improves and so does the throughput. However, the throughput declines if uninformed customers cannot be further incentivized since they have already all chosen to join. In contrast, if the offered load is high, growing information prevalence helps minimize the risk of idleness by more effectively matching capacity with demand.

Our reasoning can easily explain the findings of Chen and Frank (2004). When $\rho < \rho^*$, the offered load is low. The provider wants to increase the probability that each customer will join. Then, an unobservable queue is favorable since customers will only join under positive utility if they are informed, but will tolerate zero utility if they are uninformed. When $\rho > \rho^*$, the offered load is high. Minimizing mismatch due to uncertainty through information disclosure is an effective tactic. Thus, an observable queue is preferred.

Recall that the equilibrium strategy of uninformed customers depends on the offered load according to Theorem 3.5. Thus, the following result, as a direct corollary from Theorems 3.5 and 3.7, specifies the impact of growing information prevalence on the throughput in the primitive space (ρ, ν).

Corollary 3.8 *For given ρ and ν, define $\underline{\rho} \equiv 1 - 1/\nu$ and $\overline{\rho} \equiv y^*(\nu)$ respectively. Then,*

(i) If $0 < \rho < \overline{\rho}$, the throughput $\lambda(q^*)$ is strictly decreasing in γ.

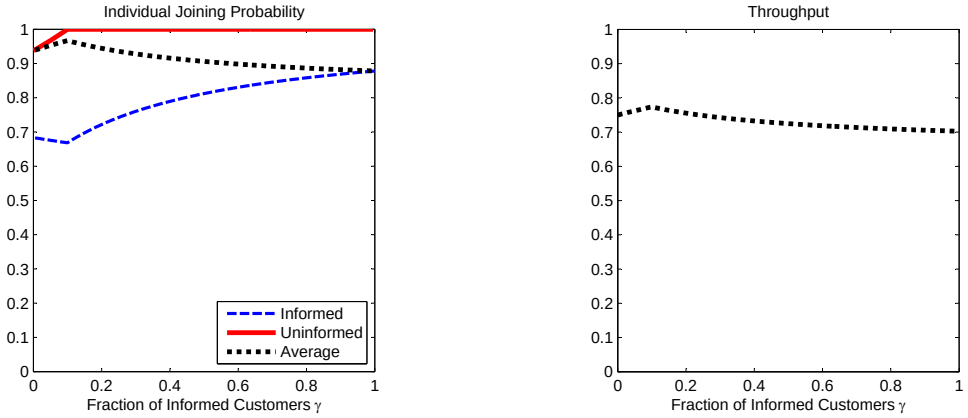
(ii) If $\overline{\rho} \leq \rho \leq \underline{\rho}$, the throughput $\lambda(q^*)$ is strictly increasing in $\gamma \in [0, \gamma_1^*(\rho, \nu))$ and is strictly decreasing in $\gamma \in [\gamma_1^*(\rho, \nu), 1]$.

(iii) If $\rho > \underline{\rho}$, the throughput $\lambda(q^*)$ is strictly increasing in γ.

We refer to Figure 3.1 to illustrate Corollary 3.8. Case (i), where information always hurts the throughput, corresponds to the area below the dashed line. Case (iii), where information always benefits the throughput, corresponds to the areas above the solid line. The more intriguing case (ii), where the equilibrium throughput $\lambda(q^*)$ is unimodal in γ,

corresponds to the intermediate area between the two lines. In this area, there is always an intermediate information level $\gamma_1^*(\rho, \nu)$ that maximizes the system throughput. Therefore, the system throughput with heterogeneous congestion information outperforms its counterparts with all customers equally informed or uninformed.

Figure 3.4: Example: $\mu = 1, R = 4, c = 1$, and $\Lambda = 0.8$



(a) Joining Probability (b) Throughput

The comparison of observable and unobservable queues in Chen and Frank (2004) is a special case of Corollary 3.8 for $\gamma = 0$ and $\gamma = 1$. Our result shows that the unobservable and observable queues are preferred in cases (i) and (iii) respectively. Therefore, Corollary 3.8 implies that the critical level ρ^* in Chen and Frank (2004) must lie between $\underline{\rho}$ and $\overline{\rho}$. For instance, Figure 3.3 shows that the observable queue is more favorable than an unobservable counterpart when $\rho = 1.1$, whereas the converse holds when $\rho = 0.8$ as shown in Figure 3.4.

In this subsection, we proved that information heterogeneity in the real-time queue length can effectively improve the system throughput, except for the case in which the offered load is sufficiently low or high. Our result generalizes the comparison of Chen and Frank (2004) to account for more realistic circumstances and also identifies the crucial role that uninformed customers play in a service system.

3.4.2 Social Welfare

In this subsection, we devote our attention to social welfare. Previous literature argues that real-time congestion information is efficient in improving social welfare because it helps customers make efficient decisions: they do not join a long queue and do not balk from a short one. Thus, it is believed that delay information should be disclosed for the

sake of society. Thanks to the advances in information technology, it is easier than ever to obtain all kinds of congestion information for public facilities, e.g., border services and highways. Is it really the case that all customers being informed always maximizes social welfare? We will answer that question in this subsection. We first discuss the influence of the increasing availability of information on individual net utility and then consider total social welfare as a whole.

For ease of exposition, we denote the individual net utility of informed and uninformed customers by $\overline{S}_{\text{I}}(q)$ and $\overline{S}_{\text{U}}(q)$, respectively. Specifically,

$$\overline{S}_{\text{I}}(q) = \sum_{i=0}^{n-1} p_i(q) \left(R - c\frac{i+1}{\mu} \right) \text{ and } \overline{S}_{\text{U}}(q) = q \sum_{i=0}^{\infty} p_i(q) \left(R - c\frac{i+1}{\mu} \right) = q(R - cW(q)).$$

By definition, an individual uninformed customer receives a non-zero utility only if $q^* = 1$ in equilibrium. An informed customer earns not only a non-negative utility but also a higher utility than an uninformed customer does at any information level. It turns out the monotonicities of $\overline{S}_{\text{I}}$ and $\overline{S}_{\text{U}}$ in the information level are also uniquely determined by the *type* of uninformed customers' equilibrium actions.

Theorem 3.9 (COMPARATIVE STATICS OF INDIVIDUAL WELFARE) *(i) $\overline{S}_{\text{I}}(q^*)$ is strictly decreasing in γ if $0 \le q^* < 1$ and is strictly increasing in γ if $q^* = 1$.*

(ii) $\overline{S}_{\text{U}}(q^*) = 0$ if $0 \le q^* < 1$ and $\overline{S}_{\text{U}}(q^*)$ is strictly increasing in γ if $q^* = 1$.

The results in the above theorem can be considered as implications from Theorem 3.7. When $0 \le q^* < 1$, the system throughput increases in the information level γ by Theorem 3.7. Hence, the system is expected to be more congested, a situation which erodes the net utility of each informed individual. Yet when $q^* = 1$, as γ increases, the throughput decreases and the system congestion is relieved. Hence, both informed and uninformed individuals obtain more net utility.

We next consider the total consumer net utility, i.e., social welfare. By (3.8), we have

$$S(q) = S_{\text{I}}(q) + S_{\text{U}}(q) = \overline{S}_{\text{I}}(q) \cdot \gamma\Lambda + \overline{S}_{\text{U}}(q) \cdot (1-\gamma)\Lambda.$$

In the case of $q^* \in [0, 1)$, since $S_{\text{U}}(q^*) = 0$, social welfare is simply the total utility of informed customers. Although the individual net utility of informed customers decreases in γ for $q^* \in [0, 1)$, the number of informed customers also increases. The next result shows that information ubiquity improves social welfare unless no uninformed customers are interested in the service, i.e., $q^* = 0$.

Theorem 3.10 (Comparative Statics of Social Welfare) *(i) If $q^* = 0$, the social welfare $S_{\text{I}}(q^*) + S_{\text{U}}(q^*)$ is strictly increasing in γ for $1 \leq \nu < 2$ and is strictly decreasing in γ for $\nu \geq 2$.*

(ii) If $0 < q^* \leq 1$, the social welfare $S_{\text{I}}(q^*) + S_{\text{U}}(q^*)$ is strictly increasing in γ.

Theorem 3.10 first confirms the conventional wisdom that in general, real-time congestion information can efficiently match the potential available capacity of a system with customer demand. That is, with congestion information, customers are able to join the service line immediately whenever they observe a short queue, which signals forthcoming availability of the server. We may therefore expect that real-time information always improves social welfare. However, a potentially negative effect due to (negative) queueing externality may also arise from the disclosure of congestion information, as Theorem 3.9(i) reveals. Growing information prevalence may also result in constantly declining individual utility, specifically when the system faces a high offered load. As the fraction of informed customers increases, more informed customers are competing with one another for the desirable positions that have fewer than n customers ahead of them. Because of the high offered load and the efficiency of information in matching waiting slots with demand, such positions are quickly taken as soon as they are available. Therefore, as γ increases, informed customers are expected to see a longer queue upon arrival and to be more likely to balk, both of which cases reduce their individual net utility.

The diminishing individual net utility of informed customers would not cause a loss of social welfare as long as uninformed customers are still interested in the service, i.e., $q^* > 0$. That is because, given their disadvantage from not being able to make instantaneous responses to system states, uninformed customers compromise by joining the queue infrequently when they expect a long line due to a high offered load. The low incentive for them to join the line helps mitigate the congestion. When an even larger portion of the high-volume customers become informed, uninformed customers have an even lower incentive to join the line. This declining interest in the service frees up the tight capacity for informed customers, who earn higher utility than uninformed customers, who earn zero utility. More importantly, the reduced participation of uninformed customers alleviates competition among informed customers for the appealing positions and thus prevents the informed customers' joining probability and individual net utility from descending quickly. Nonetheless, after all uninformed customers lose their interest in the service and choose to balk, i.e., $q^* = 0$, more prevalent information only increases the number of informed customers. Without the uninformed customers' compromise, the increased number of informed customers significantly intensifies the competition among

informed customers for the appealing queue positions. The individual net utility of an informed customer therefore decreases faster when $q^* = 0$ than when $q^* > 0$. As a result, if no uninformed customers consider joining, social welfare starts to deteriorate when real-time delay information is more prevalent.

Figure 3.5: Example: $\mu = 1, R = 4, c = 1$, and $\Lambda = 2.3$

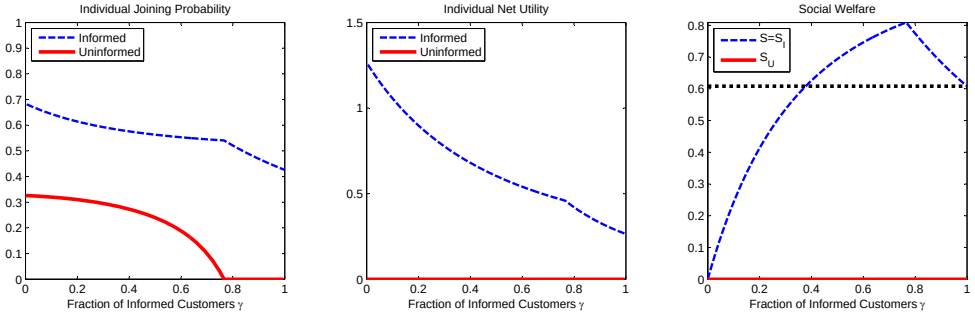


Figure 3.5, which uses $\rho = 2.3$ as an example, illustrates the patterns. For informed customers, their joining probability and individual net utility both decline with the information level γ. However, when $q^* > 0$, both metrics decease noticeably more slowly than when $q^* = 0$. As we have discussed, this difference can be explained by uninformed customers' disincentives to join. This declining incentive to join, together with the declining number of the uninformed customers, helps the system use its tight capacity to serve more customers who can yield higher welfare. These effects no longer exist when the uninformed customers completely refrain from joining.

Combining Theorems 3.5 and 3.10, we can conclude what the monotonicity properties of the social welfare are under different offered loads.

Corollary 3.11 *For $1 \le \nu < 2$, the social welfare is strictly increasing in γ. For $\nu \ge 2$,*

(i) If $0 \le \rho \le y^*(\nu)$, the equilibrium social welfare is strictly increasing in γ.

(ii) If $\rho > y^*(\nu)$, the equilibrium social welfare is strictly increasing in $\gamma \in [0, \gamma_0^*(\rho, \nu))$ and is decreasing in $\gamma \in [\gamma_0^*(\rho, \nu), 1]$.

Theorem 3.10 and Corollary 3.11 have two implications. First, absolute transparency in congestion information may not achieve the most economic efficiency. Specifically, a highly loaded system yields the most social welfare if some customers are uninformed. Second, even if information is not at the optimal level $\gamma_0^*(\rho, \nu)$, the social welfare under heterogeneous information can still outperform a completely observable counterpart for

a large range of information levels. For instance, social welfare for any $\gamma \in (0.38, 1)$ is higher than when $\gamma = 1$ in the example displayed by Figure 3.5.

As an exception, if customers have a very low service reward R such that the informed ones join the queue only when it is empty, in which case $1 \leq \nu < 2$, information always helps, even when $q^* = 0$. From the perspective of a server, the line temporarily holds waiting customers and supplies the server with customers upon completion of a job. If $q^* = 0$, even uninformed customers do not join a queue. As a result, no one intends to queue at all. After completing a service request, the server has to remain idle and wait until the next customer arrives to resume generating welfare. Therefore, the server has a strong incentive to rely on the ubiquity of information to secure a customer as soon as its capacity is available. Such different behavior from the general case, when informed customers are willing to be the only one in the queue, is also observed in Hassin (1986).

Summary. Table 3.1 summarizes the effects of growing information prevalence on various performance measures. The effects depend on the type of the equilibrium joining strategy by uninformed customers, specified by their equilibrium joining probability q^*. One can easily visualize the effects of information prevalence and the optimal information levels in the primitive space (ρ, ν), illustrated in Figure 3.1, by combining Theorem 3.5 and Table 3.1.

Table 3.1: Comparative Statics in γ

	$q^* = 0$	$0 < q^* < 1$	$q^* = 1$
Throughput	$\uparrow$	$\uparrow$	$\downarrow$
Individual welfare of informed customers	$\downarrow$	$\downarrow$	$\uparrow$
Individual welfare of uninformed customers	0	0	$\uparrow$
Social welfare	$\downarrow^*$	$\uparrow$	$\uparrow$

$*$: $\uparrow$ for $\nu \in [1, 2)$.

A central question in any resource allocation problem is to define the notion of efficiency. To an engineer, throughput, which implies the utilization of the server, is the relevant efficiency measure. To an economist, social welfare, which focuses on the overall economic benefit, may be a more suitable measure. We show that the effects of growing information prevalence on these two efficiency measures do not go in the same direction in the two extreme cases $q^* = 0$ and $q^* = 1$.

If throughput is the focal performance measure, the service provider should reveal the queue length and encourage information dissemination when the offered load is high enough (the top area in Figure 3.1) and conceal it when the offered load is low enough (the bottom area in Figure 3.1). Otherwise, it is optimal to have a segment of uninformed

customers or reveal the real-time information only to a fraction of customers if the offered load is in an intermediate range (the middle area in Figure 3.1).

If social welfare is the focal performance measure, the service provider should reveal the queue length information and encourage its dissemination when the offered load is relatively small (the areas below the solid line in Figure 3.1). In other situations (the area above the solid line, i.e., the top area, in Figure 3.1), it is optimal to have a segment of uninformed customers or, equivalently, the real-time congestion information should be hidden from certain customers.

3.5 Endogenizing Information Levels

Our base model provides an optimistic view on the impacts of information heterogeneity. The fact that some customers do not possess real-time delay information indeed helps the system throughput when the offered load is modest and improves the social welfare when the offered load is high.

However, it should be noted that information ignorance would normally be considered *irrational* if access to the information is completely free. Uninformed customers who do not obtain real-time information always earn less utility than informed customers. Therefore, if congestion information is free and convenient, a rational, uninformed customer has every incentive to learn the queue length. Consequently, all customers would choose to be informed in equilibrium under self-interested rational choices. Next we discuss how the service provider can achieve the optimal information level by charging an information fee, when customers are completely rational. For this purpose, we temporarily assume in this section that all customers exhibit self-interested, utility-maximization behavior and they know the system parameters Λ, μ, R and c.

3.5.1 Inducing the Optimal Throughput

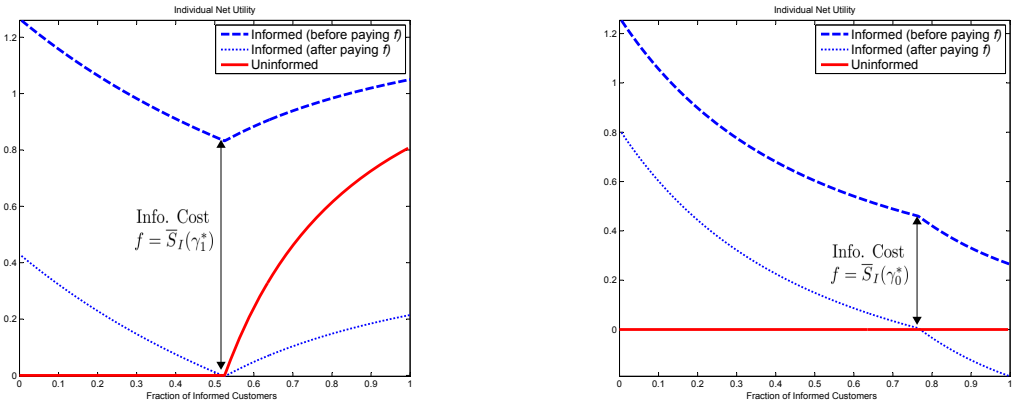
The optimal information levels can be easily induced when the offered load is either very high or low. Recall that if $\rho > \overline{\rho}$, the system throughput is maximized when all customers are informed; i.e., $\gamma = 1$. Therefore, with congestion information revealed by the provider, the self-interested choice by all customers to be informed leads to an equilibrium that is also sustained as a system optimum. If $\rho < \underline{\rho}$, no one's being informed leads to the maximum throughput. Thus, the provider can simply conceal real-time congestion information such that no customers can be informed.

In the case where the offered load is in an intermediate range, i.e., when $\underline{\rho} \leq \rho \leq \overline{\rho}$, the

system throughput is maximized at $\gamma = \gamma_1^*$ according to Corollary 3.8. In this case, the optimal information level is not achievable through decentralized actions by customers under either queue-length transparency or secrecy. An information fee can resolve this issue. As long as the exact fraction γ_1^* of customers are induced to pay the fee, the throughput is maximized.

Proposition 3.12 (INFORMATION FEE FOR OPTIMAL THROUGHPUT) *Assume customers are rational. If the offered load $\rho \in [\underline{\rho}, \overline{\rho}]$, the service provider induces the optimal information level that maximizes throughput, by charging an information fee $f = \overline{S}_{\text{I}}(\gamma_1^*)$.*

Figure 3.6: **Illustration of Information Fee** ($\mu = 1, R = 4$, **and** $c = 1$)



(a) $\rho = \Lambda/\mu = 1.1$ (b) $\rho = \Lambda/\mu = 2.3$

We use Figure 3.6(a) where $\rho = 1.1$ to illustrate Proposition 3.12. We first show that for the base model, if an exogenous fraction of informed customers are forced to pay an information fee $f = \overline{S}_{\text{I}}(\gamma_1^*)$ and the rest stay uninformed, the informed customers still earn a non-negative utility after paying the fee and the system behaves exactly the same as in the base model. With no information fee, being informed yields higher individual net utility than being uninformed, as the bold dashed curve stays above the solid one in Figure 3.6(a). Note that the individual net utility of an informed customer reaches its minimum when $\gamma = \gamma_1^* = 0.51$, at which point the system throughput peaks (see Figure 3.3). Now, assume that informed customers are forced to pay a fee $f = \overline{S}_{\text{I}}(\gamma_1^*) - \overline{S}_{\text{U}}(\gamma_1^*) = \overline{S}_{\text{I}}(\gamma_1^*)$, which equals the utility difference of an informed customer and an uninformed one at $\gamma = \gamma_1^*$, for inspecting the queue length. Then the utility curve of informed customers shifts downward to the position displayed by the thin dotted line in Figure 3.6(a). As shown in the plot, informed customers still earn a non-negative utility after paying the

information fee, regardless of their fraction in the population. This implies that the additional information fee would not alter the joining behavior of informed customers, if they have to pay the fee. Hence, the irrational, uninformed customers have the same equilibrium joining probability q^* as before, regardless of the size of informed customers. In other words, if informed customers are required to pay the fee $f = \overline{S}_{\text{I}}(\gamma_1^*)$, the system behaves in exactly the same way as in the base model, except that informed customers receive less utility at any information level γ.

Given the characterized system dynamics with the information fee $f = \overline{S}_{\text{I}}(\gamma_1^*)$ for any given γ, we further show that in the current setting under self-interested choices by all customers, the system reaches an equilibrium in which γ_1^* fraction of customers are willing to pay the fee to be informed and the rest are not willing to do so and hence stay uninformed. Suppose that the system is in a state where less than γ_1^* fraction pay the fee and are informed; i.e., $\gamma < \gamma_1^*$. From Figure 3.6(a), we see that not paying the fee and staying uninformed earns zero utility and is dominated by paying the information fee and being informed; i.e., $\overline{S}_{\text{I}}(\gamma) - f > \overline{S}_{\text{U}}(\gamma) = 0$ for $\gamma < \gamma_1^*$. Therefore, some uninformed customers have the incentive to become informed until $\gamma = \gamma_1^*$. Next, consider the system in a state where more than γ_1^* fraction are informed; i.e., $\gamma > \gamma_1^*$. In this case, paying the information fee is not worthwhile. Saving the cost and being uninformed yields higher utility; i.e., $\overline{S}_{\text{I}}(\gamma) - f < \overline{S}_{\text{U}}(\gamma)$ for $\gamma > \gamma_1^*$. Therefore, some informed customers have the incentive not to pay for real-time information and become uninformed until $\gamma = \gamma_1^*$. As a result, at $\gamma = \gamma_1^*$, both options, being informed or uninformed, are equally appealing. The system reaches an equilibrium, in which the information level that maximizes the throughput is induced through customer decentralized choices under the information fee $f = \overline{S}_{\text{I}}(\gamma_1^*)$.

3.5.2 Inducing the Optimal Social Welfare

According to Corollary 3.11, if the offered load is not high, i.e., if $\rho \leq \overline{\rho}$, the system attains its optimal social welfare when all customers are informed. This can be accomplished by customer self-interested choices under queue-length information transparency, since being informed is a dominant strategy. Hence, social welfare optimality can be achieved without any coercion as long as the service provider reveals the real-time congestion information. In contrast, if the offered load is high, i.e., if $\rho > \overline{\rho}$, the optimal social welfare is achieved when $\gamma = \gamma_0^*$, at which point uninformed customers have an incentive to become informed. The social welfare optimality thus cannot be established through decentralized decisions under information transparency. The service provider has to

charge an information fee to achieve the socially optimal solution.

Notice that with an information fee, the social welfare includes total customer welfare and service provider's collected fees (cf. Hassin and Haviv 2003 p. 49); i.e.,

$$S = \left(\sum_{i=0}^{n-1} p_i(q) \left(R - c\frac{i+1}{\mu} \right) - f \right) \cdot \gamma\Lambda + q \cdot (R - cW(q)) \cdot (1-\gamma)\Lambda + f \cdot \gamma\Lambda.$$

The term $f \cdot \gamma\Lambda$ cancels out, reflecting the fact that from the perspective of the entire society, the information fee is only a transfer payment that has no effect on the value of social welfare itself but can help regulate the demand side and potentially achieve social optimality.

Proposition 3.13 (Information Fee for Optimal Social Welfare) *Assume customers are rational. If the offered load $\rho > \overline{\rho}$, the service provider induces the optimal information level that maximizes the social welfare, by charging an information fee $f = \overline{S}_{\text{I}}(\gamma_0^*)$.*

The idea behind Proposition 3.13 is similar to that behind Proposition 3.12, but with a minor difference. We use Figure 3.6(b) where $\rho = 2.3$ for illustration. As in the example displayed by Figure 3.6(a), if all informed customers are required to pay the information fee (assuming they cannot choose not to pay the fee and become uninformed), their individual utility curve shifts from the bold dashed line to the thin dotted one. For any $\gamma < \gamma_0^*$, informed customers still earn positive utility after information payment, which is more than the zero utility of uninformed customers. Hence, some uninformed customers would like to inspect the queue by paying the fee. The incentive for converting from the uninformed to the informed vanishes until $\gamma = \gamma_0^*$, at which point being informed or uninformed receives the same individual net utility. So far, the rationale for Proposition 3.13 is the same as that for Proposition 3.12. The difference comes when $\gamma > \gamma_0^*$. As illustrated by the thin dotted line in Figure 3.6(b), for $\gamma > \gamma_0^*$, the γ fraction of informed customers would incur negative utility after paying the information fee. This implies that under self-interested choices, paying an information fee $f = \overline{S}_{\text{I}}(\gamma_0^*)$ is not individually rational for the informed customers whose fraction is more than γ_0^*. As a result, $\gamma = \gamma_0^*$ emerges as an equilibrium through customer decentralized decisions when the service provider charges an information fee $f = \overline{S}_{\text{I}}(\gamma_0^*)$ for inspecting the queue.

Although our discussion focuses on achieving the optimal information level to maximize throughput or social welfare, the service provider may have other objectives and can charge a different information fee to achieve another desired information level.

3.6 Impacts of Heterogeneous Customer Characteristics

In the base model, we treated all customers as identical agents except for their possession of real-time congestion information. It is plausible that informed customers might have other characteristics different from the uninformed, which also could explain the difference in their information possession. For instance, informed customers tend to own a smart phone, be more technology savvy and younger, and hence might be less patient. In this section, we explore how other heterogeneities, in addition to awareness of real-time congestion, in customer characteristics may interact with information heterogeneity in affecting the throughput and social welfare. Specifically, we assume that informed and uninformed customers receive a reward of, respectively, R_{I} and R_{U} from the service. Moreover, their respective unit waiting costs are c_{I} and c_{U}. In terms of joining strategy, informed customers still use a threshold policy: They join the line if and only if the queue length is less than $n_{\text{I}} \equiv \lfloor \nu_{\text{I}} \rfloor \equiv \lfloor R_{\text{I}}\mu/c_{\text{I}} \rfloor$; otherwise, they choose to balk. In contrast, uninformed customers choose their joining probability q according to the expected utility $R_{\text{U}} - c_{\text{U}}W(q)$. In other words, the dynamics of this extended model evolve in the same vein as the base model. As shown below, the additional customer heterogeneities do *not* change our result for the system throughput.

Theorem 3.14 (Comparative Statics of Throughput) *Consider the model with heterogeneous customer characteristics.*

(i) If $0 \le q^* < 1$, the throughput $\lambda(q^*)$ is strictly increasing in γ.

(ii) If $q^* = 1$, the throughput $\lambda(q^*)$ is strictly decreasing in γ.

Recall that maximizing the system throughput is equivalent to minimizing the idleness of the server, which results from inadequate service requests and an intertemporal mismatch between capacity and demand, i.e., the mean effect and the variability effect. As we have argued, ubiquity of congestion information improves throughput by overcoming the mean effect under low offered loads by incentivizing uninformed customers, and effectively reducing the variability effect under high offered loads by matching capacity and waiting slots better with informed customers intertemporally. The influence of information remains robust when service rewards and unit waiting costs become heterogeneous. So is our result for throughput.

On the other hand, social welfare directly measures the total service rewards less the costs of delay. Thus, the result for social welfare is expected to be affected by customer

heterogeneities. The following result indicates the key conditions that might cause a difference.

Theorem 3.15 (Comparative Statics of Social Welfare) *Consider the model with heterogeneous customer characteristics. Let* $\nu_{\text{I}} \equiv R_{\text{I}}\mu/c_{\text{I}} \geq 2$ *and* $\nu_{\text{U}} \equiv R_{\text{U}}\mu/c_{\text{U}}$.

(i) If $q^* = 0$, *the social welfare* $S_{\text{I}}(q^*) + S_{\text{U}}(q^*)$ *is strictly decreasing in* γ.

(ii) If $0 < q^* < 1$, *the social welfare* $S_{\text{I}}(q^*) + S_{\text{U}}(q^*)$ *is strictly increasing in* γ *if* $\nu_{\text{I}} \geq \nu_{\text{U}}$. *Otherwise, the social welfare might be unimodal in* γ.

(iii) If $q^* = 1$, *the social welfare* $S_{\text{I}}(q^*) + S_{\text{U}}(q^*)$ *is strictly increasing in* γ *if* $\lfloor \nu_{\text{I}} \rfloor \geq \nu_{\text{U}} - \frac{1+(1-\gamma)\rho}{1-(1-\gamma)\rho}$. *Otherwise, the social welfare might be unimodal in* γ.

Theorem 3.15 reveals that if ν_{I} is no less than ν_{U}, our welfare result for the homogeneous case still holds for the heterogeneous extension. Otherwise, the social welfare can even be unimodal in the information level over the range such that $q^* \in (0, 1)$ or $q^* = 1$.

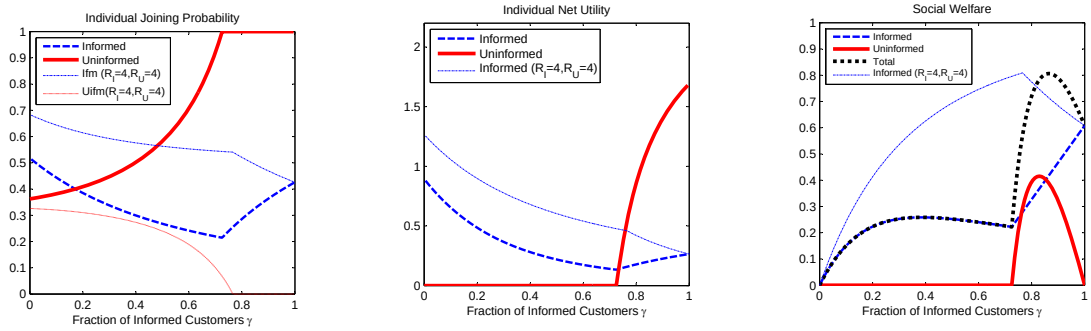
The parameters ν_{I} and ν_{U} are the joining thresholds for informed and uninformed customers if they can observe the queue length. Since customers who tolerate a longer queue implies their higher valuation of service relative to their unit waiting cost, we can consider ν_{I} and ν_{U} as normalized service valuations of each customer segment. In the homogeneous case, in which both segments of customers value the service equally, the monotonicity of the social welfare in the information level γ is primarily determined by the total utility of the informed customers. Such an effect is expected to be more salient if informed customers value the service more than the uninformed ones. Thus, the result in the heterogeneous case is similar to the homogeneous case if ν_{I} is relatively larger than ν_{U} as stated in Theorem 3.15, parts (ii) and (iii).

On the contrary, if uninformed customers relatively value the service more, they have a stronger incentive to seek service, and that may cause the monotonicity property of the social welfare to be different from that of the homogeneous case. We use an example, which is displayed in Figure 3.7, to illustrate the differences in uninformed customers' equilibrium behavior and their impacts on informed customers and social welfare. For comparison with the homogeneous case, we choose the same system parameters as those in Figure 3.5, except that uninformed customers receive a reward $R_{\text{U}} = 6$ instead of 4 while the reward for the informed, R_{I}, is still 4.

We first discuss the differences in the incentive and behavior of uninformed customers between the heterogeneous and homogeneous cases, and its impact on the informed customers whose joining threshold remains unchanged. For the same information level γ,

uninformed customers are more enthusiastic about joining the queue in the heterogeneous case than they are in the homogeneous case.[7] This is because uninformed customers value the service higher than they do in the homogeneous case. Then they earn higher individual net utility for any given queue length. Hence, their equilibrium joining probability q^*, represented by the solid curve in Figure 3.7(a), is higher than it is in the homogeneous case, represented by the dotted curve in the same plot. Such an increased joining probability leads to more uninformed customers competing with informed ones for waiting positions with fewer than 4 people ahead. As a result, informed customers are less likely to join the line and earn positive utilities in the heterogeneous case than in the homogeneous case for the same information level γ. This explains why the dashed lines stay below the dash-dot ones in Figures 3.7(a) and 3.7(b) , respectively. Furthermore, due to larger externalities exerted by uninformed ones (as a larger q^* at the same γ) in the heterogeneous case, the total welfare generated by the informed segment is lower than it is in the homogeneous case, as displayed by the bold dashed curve and the thin dotted one in Figure 3.7(c).

Figure 3.7: Example: $\mu = 1, R_{\text{I}} = 4, R_{\text{U}} = 6, c = 1$, and $\Lambda = 2.3$



(a) Joining Probability (b) Individual Net Utility (c) Social Welfare

For the heterogeneous case, uninformed customers' equilibrium joining probability q^* increases in the information level γ, whereas the joining probability of informed customers first declines then increases in γ. In contrast, for the homogeneous case, q^* decreases in the information level γ given the relatively high offered load $\rho = 2.3$. Since $\nu_{\text{U}} = 6 > \nu_{\text{I}} = 4$, 4 or 5 customers waiting ahead is only welcome by uninformed customers but is not acceptable to informed ones. As γ increases, the number of uninformed customers declines. Therefore, the fourth and fifth positions in the queue are more likely to be available, and that contributes to an increasing equilibrium joining probability q^* by

[7] In fact, it can be rigorously proved that if $\nu_{\text{U}} \geq \lfloor \nu_{\text{I}} \rfloor + 1$, uninformed customers will never choose to always balk in equilibrium regardless of the offered load $\rho = \Lambda/\mu$. This is not the case if $\nu_{\text{U}} = \nu_{\text{I}}$.

the uninformed customers. While q^* increases in γ, the probability that an informed customer observes a queue shorter than n_{I} decreases. This decay stops when q^* reaches 1, corresponding to $\gamma = 0.72$, beyond which a further increase in γ only lowers the number of uninformed customers with no possibility of boosting q^* further. Therefore, congestion starts to be alleviated as the throughput starts to decrease (see Theorem 3.14(ii)). As a result, the joining probability of informed customers increases in γ after q^* hits 1. So do the individual net utilities of informed and uninformed customers, as shown in the second plot of Figure 3.7(b).[8]

The increasing joining probability q^* of uninformed customers in the information level γ can result in the unimodal behavior of social welfare over the range of $q^* \in (0, 1)$. As γ grows, uninformed customers join the queue more often, and thus inflict more negative externalities on informed customers in the heterogenous case. In contrast, as γ increases, uninformed customers join the queue less often in the homogeneous case. As a result, in the heterogeneous case, the consumer welfare of the informed segment, which also equals the social welfare, does not necessarily increase in the growing information prevalence like the homogeneous case, and it may decline from a certain point, as shown in Figure 3.7(c).

When $q^* = 1$, the number of informed customers and their individual net utility both increase in γ. The total consumer welfare of the informed customers hence must increase in γ. For the uninformed segment, although the net utility of an uninformed customer increases from zero, the size of the segment shrinks to zero. The total uninformed customer welfare is thus unimodal in γ. In the heterogeneous case, the welfare of the uninformed customers contributes to a larger portion of social welfare. Thus, the unimodal behavior of the uninformed customers' welfare can lead to the unimodal behavior of social welfare in γ over the range of $q^* = 1$.

In summary, when customers of different information segments exhibit different characteristics, the results for the system throughput remain the same as in the homogeneous case. However, the results for social welfare can be different. As a result, in addition to the effect of information heterogeneity, the non-monotonic behavior of social welfare in the information level may also result from uninformed customers' high service valuation or low unit delay cost. In reality, uninformed customers could also have different characteristics, for which case our numerical experiments confirm similar insights.

[8]Note that since $R_{\text{U}} > R_{\text{I}}$, an uninformed customer is able to receive higher individual utility than an informed customer when $q^* = 1$, as shown in Figure 3.7(b). This situation never occurs in the homogeneous case.

3.7 Conclusion

We have considered information heterogeneity in a service system and described the effect of more informed customers in the population on various performance measures. In particular, that effect can be determined by the type of the equilibrium joining behavior of uninformed customers. Perhaps surprisingly, we have shown that more informed customers may not necessarily benefit throughput or social welfare.

Our results suggest that the presence of uninformed customers who interact with informed customers does not necessarily jeopardize system performance. The information ignorance behavior may not be as detrimental as one might expect. In fact, the information heterogeneity helps the system in certain conditions. Our results may raise the question whether the current practice of disseminating free delay information is the most effective approach to managing congestion and whether an information fee might be introduced to intentionally create heterogeneity in the possession of delay information. Another implication of our results is that service providers can be better off by limiting access to real-time information about delays so as to intentionally create a mix of informed and uninformed customers. Our findings may well justify Disney World's practice of allowing only premium customers to obtain waiting-time information about its popular attractions.[9] In that case, the amusement park may try to maximize the total satisfaction levels associated with customer experiences in the park. Another possible way of controlling the availability of congestion information is through targeted delay announcements. For a loaded call center, the service provider may consider making delay announcements only to a fraction of callers, e.g., loyal customers.

3.8 Appendix

3.8.1 Technical Results

Lemma 3.16 *The function $L(\rho, \nu)$ defined in Corollary 3.4 is strictly decreasing in ρ.*

Proof of Lemma 3.16. For notation simplicity, we suppress $L(\rho, \nu)$'s dependence on ν and write $L(\rho)$ or simply L. By the definition of $L(\rho)$,

$$\frac{dL}{d\rho} = \frac{\phi(\rho)}{\rho^{n+1}(\rho - 1)^3},$$

[9] A `TouringPlans.com` Premium Subscription is needed for access to the data on waiting times and crowds through a mobile app.

where

$$\phi(\rho) \equiv \nu(n+1)\rho^2 + (2-\nu-2n\nu+n)\rho + n\nu - n - \langle\nu\rangle\rho^{n+2} + (\langle\nu\rangle - 2)\rho^{n+1}.$$

Taking first and second derivatives of $\phi(\rho)$ with respect to ρ, we have

$$\phi'(\rho) = \frac{d\phi}{d\rho} = 2\nu(n+1)\rho + (2-\nu-2n\nu+n) - (n+2)\langle\nu\rangle\rho^{n+1} + (n+1)(\langle\nu\rangle - 2)\rho^n \quad (3.9)$$

and

$$\begin{aligned}
\phi''(\rho) = \frac{d^2\phi}{d\rho^2} &= 2\nu(n+1) - (n+1)(n+2)\langle\nu\rangle\rho^n + n(n+1)(\langle\nu\rangle - 2)\rho^{n-1} \\
&= (n+1)\left[2\nu - 2n\rho^{n-1} - \langle\nu\rangle\left(2\rho^n + n\rho^n - n\rho^{n-1}\right)\right] \\
&\overset{\nu=n+\langle\nu\rangle}{=} (n+1)\left[2n(1-\rho^{n-1}) + \langle\nu\rangle\left(2(1-\rho^n) + n\rho^{n-1}(1-\rho)\right)\right] \\
&= (n+1)(1-\rho)\left[2n\sum_{i=0}^{n-2}\rho^i + \langle\nu\rangle\left(2\sum_{i=0}^{n-1}\rho^i + n\rho^{n-1}\right)\right], \quad (3.10)
\end{aligned}$$

where $\sum_{i=0}^{n-2}\rho^i$ is understood as 0 for $n=1$. Moreover, note that $\phi(1) = \phi'(1) = 0$. Hence, by Eq.(3.9) and (3.10),

$$\begin{cases} \phi''(\rho) > 0, & \text{if } 0<\rho<1 \\ \phi''(\rho) < 0, & \text{if } \rho > 1 \end{cases} \Longrightarrow \begin{cases} \phi'(\rho) < \phi'(1) = 0, & \text{if } 0<\rho<1 \\ \phi'(\rho) < \phi'(1) = 0, & \text{if } \rho > 1 \end{cases} \Longrightarrow \begin{cases} \phi(\rho) > \phi(1) = 0, & \text{if } 0<\rho<1 \\ \phi(\rho) < \phi(1) = 0, & \text{if } \rho > 1 \end{cases}.$$

Therefore, $\dfrac{dL}{d\rho} = \dfrac{\phi(\rho)}{\rho^{n+1}(\rho-1)^3} < 0$ for $0<\rho<1$ and $\rho>1$. Finally, by L'Hôpital's rule, $\lim\limits_{\rho\to 1}\dfrac{dL}{d\rho} = n(n+1)(n+2-3\nu)/6$, which is negative for all $\nu > 1$ and is zero for $\nu = 1$. We thus conclude that $\dfrac{dL}{d\rho} < 0$ for $\rho > 0$ (almost surely except for the point $\rho = 1$ when $\nu = 1$), i.e., $L(\rho)$ is strictly decreasing in ρ (note that the derivative being equal to 0 at one point does not affect the strict monotonicity of a function). ■

Lemma 3.17 *In the neighborhood where full participation is not adopted by uninformed customers in equilibrium, i.e., $q^* \in [0,1)$, for any information level γ', there exists $k < n$ such that*

$$\left.\frac{dp_i(q^*(\gamma))}{d\gamma}\right|_{\gamma=\gamma'} < 0 \text{ for } 0 \le i \le k \quad \text{and} \quad \left.\frac{dp_i(q^*(\gamma))}{d\gamma}\right|_{\gamma=\gamma'} \ge 0 \text{ for } k < i < n.$$

Proof of Lemma 3.17. We have shown, in Lemma 3.6, that $p_0(q^*(\gamma))$ strictly decreases in γ for $0 \le q^* < 1$. At $\gamma = \gamma'$, if for any $i = 1, \ldots, n-1$, $dp_i(q^*(\gamma'))/d\gamma < 0$. Then,

$k = n-1$.

If there exists $k < n-1$, such that $dp_k(q^*(\gamma'))/d\gamma \geq 0$ at γ', then the statement holds as long as for any $i = k, k+1, \ldots, n-1$, $dp_i(q^*(\gamma'))d\gamma \geq 0$. Let $\rho_c(\gamma) = \gamma\rho + q^*(\gamma)(1-\gamma)\rho$, where $0 \leq q^*(\gamma) < 1$. By Eq. (3.3), $p_i(q^*(\gamma)) = p_k(q^*(\gamma))\rho_c^{i-k}(\gamma) = p_0(q^*(\gamma))\rho_c^k(\gamma)\rho_c^{i-k}(\gamma)$. Thereby, for $i = k, k+1, \ldots, n-1$,

$$\frac{dp_i(q^*(\gamma))}{d\gamma} = \frac{dp_k(q^*(\gamma))}{d\gamma}\underbrace{\rho_c^{i-k}(\gamma)}_{\geq 0} + \underbrace{p_k(q^*(\gamma))(i-k)\rho_u^{i-k-1}(\gamma)}_{\geq 0}\frac{d\rho_c(\gamma)}{d\gamma}.$$

At γ', $dp_k(q^*(\gamma'))/d\gamma \geq 0$ by assumption. Hence, if $d\rho_c(\gamma')/d\gamma \geq 0$, $dp_i(q^*(\gamma'))/d\gamma \geq 0$. Note that

$$\frac{dp_k(q^*(\gamma))}{d\gamma} = \frac{dp_0(q^*(\gamma))}{d\gamma}\rho_c^k(\gamma) + p_0(q^*(\gamma))k\rho_c^{k-1}(\gamma)\frac{d\rho_c(\gamma)}{d\gamma} \geq 0.$$

The first term is negative since $p_0(q^*(\gamma))$ strictly decreases in γ. Hence, $dp_k(q^*(\gamma'))/d\gamma \geq 0$ implies $d\rho_c(\gamma')/d\gamma \geq 0$, which further leads to $dp_i(q^*(\gamma'))/d\gamma \geq 0$ for $i = k, k+1, \ldots, n-1$. ■

Proposition 3.18 (JOINING PROBABILITY FOR INFORMED CUSTOMERS)

(i) If $0 \leq q^* < 1$, the probability $\sum_{i=0}^{n-1} p_i(q^*)$ that an informed customer joins the queue is strictly decreasing in γ.

(ii) If $q^* = 1$, the probability $\sum_{i=0}^{n-1} p_i(q^*)$ that an informed customer joins the queue is strictly increasing in γ.

Proof of Proposition 3.18. (i) When $q^* = 0$, $\sum_{i=0}^{n-1} p_i(q^*=0) = p_0(q^*=0)\sum_{i=0}^{n-1}(\gamma\rho)^i = \frac{1-(\gamma\rho)^n}{1-(\gamma\rho)^{n+1}}$. It is straightforward to verify that $\frac{1-(\gamma\rho)^n}{1-(\gamma\rho)^{n+1}}, n \geq 1$ is strictly decreasing in γ.

Consider the case where $0 < q^* < 1$. Again, let $\rho_c = \rho(\gamma + q^*(1-\gamma))$. Recall that we have shown $d\rho_c/d\gamma > 0$ in the proof of Theorem 3.7. If $\sum_{i=0}^{n-1} p_i(q^*)$ is strictly decreasing in ρ_c, by the chain rule, it must be strictly decreasing in γ. Hence, it is sufficient to prove that $\sum_{i=0}^{n-1} p_i(q^*)$ is strictly decreasing in ρ_c. We rewrite

$$\begin{aligned}\sum_{i=0}^{n-1} p_i(q^*) &= \sum_{i=0}^{n-1} p_0(q^*)\rho_c^i \\ &= p_0(q^*)\frac{1-\rho_c^n}{1-\rho_c} \\ &\overset{(3.8.2)}{=} \frac{1-\rho_c^n}{1-\rho_c}\Big/\left(\frac{1-\rho_c^n}{1-\rho_c} + \frac{\rho_c^n}{1-\rho_c+\gamma\rho}\right)\end{aligned}$$

$$= \left(1+\frac{\rho_{\mathrm{c}}^n}{1-\rho_{\mathrm{c}}+\gamma\rho}\cdot\frac{1-\rho_{\mathrm{c}}}{1-\rho_{\mathrm{c}}^n}\right)^{-1}.$$

By Eq.(3.21),

$$\begin{aligned}
\frac{\rho_{\mathrm{c}}^n}{1-\rho_{\mathrm{c}}+\gamma\rho}\cdot\frac{1-\rho_{\mathrm{c}}}{1-\rho_{\mathrm{c}}^n} &= \frac{\rho_{\mathrm{c}}^n(1-\rho_{\mathrm{c}})}{1-\rho_{\mathrm{c}}^n}\cdot\frac{1}{2}\left(\langle\nu\rangle+\sqrt{\langle\nu\rangle^2+4L(\rho_{\mathrm{c}})}\right)\\
&= \frac{1}{2}\langle\nu\rangle\frac{\rho_{\mathrm{c}}^n(1-\rho_{\mathrm{c}})}{1-\rho_{\mathrm{c}}^n}+\sqrt{\left(\frac{1}{2}\langle\nu\rangle\frac{\rho_{\mathrm{c}}^n(1-\rho_{\mathrm{c}})}{1-\rho_{\mathrm{c}}^n}\right)^2+\frac{\rho_{\mathrm{c}}^{2n}(1-\rho_{\mathrm{c}})^2}{(1-\rho_{\mathrm{c}}^n)^2}L(\rho_{\mathrm{c}})}\\
&\overset{(3.20)}{=} \frac{1}{2}\langle\nu\rangle\frac{\rho_{\mathrm{c}}^n(1-\rho_{\mathrm{c}})}{1-\rho_{\mathrm{c}}^n}+\sqrt{\left(\frac{1}{2}\langle\nu\rangle\frac{\rho_{\mathrm{c}}^n(1-\rho_{\mathrm{c}})}{1-\rho_{\mathrm{c}}^n}\right)^2+\frac{\rho_{\mathrm{c}}^{2n}(1-\rho_{\mathrm{c}})^2}{(1-\rho_{\mathrm{c}}^n)^2}\cdot\frac{\langle\nu\rangle(\rho_{\mathrm{c}}-1)\rho_{\mathrm{c}}^n+\nu-\nu\rho_{\mathrm{c}}+\rho_{\mathrm{c}}^n-1}{(1-\rho_{\mathrm{c}})^2\rho_{\mathrm{c}}^n}}\\
&= \frac{1}{2}\langle\nu\rangle\frac{\rho_{\mathrm{c}}^n(1-\rho_{\mathrm{c}})}{1-\rho_{\mathrm{c}}^n}+\sqrt{\left(\frac{1}{2}\langle\nu\rangle\frac{\rho_{\mathrm{c}}^n(1-\rho_{\mathrm{c}})}{1-\rho_{\mathrm{c}}^n}\right)^2+\langle\nu\rangle\frac{\rho_{\mathrm{c}}^n(1-\rho_{\mathrm{c}})}{1-\rho_{\mathrm{c}}^n}+\rho_{\mathrm{c}}^n\frac{n-n\rho_{\mathrm{c}}+\rho_{\mathrm{c}}^n-1}{(1-\rho_{\mathrm{c}}^n)^2}}.
\end{aligned}$$

It is apparent that $\dfrac{\rho_{\mathrm{c}}^n(1-\rho_{\mathrm{c}})}{1-\rho_{\mathrm{c}}^n}=\left(\sum_{i=1}^n\rho_{\mathrm{c}}^{-i}\right)^{-1}$ is strictly increasing in ρ_{c}. Therefore, to show $\sum_{i=0}^{n-1}p_i(q^*)$ is strictly decreasing in ρ_{c}, it suffices to justify $\rho_{\mathrm{c}}^n\dfrac{n-n\rho_{\mathrm{c}}+\rho_{\mathrm{c}}^n-1}{(1-\rho_{\mathrm{c}}^n)^2}$ increases in ρ_{c}.

$$\left(\rho_{\mathrm{c}}^n\frac{n-n\rho_{\mathrm{c}}+\rho_{\mathrm{c}}^n-1}{(1-\rho_{\mathrm{c}}^n)^2}\right)'=\frac{n\rho_{\mathrm{c}}^{n-1}}{(1-\rho_{\mathrm{c}}^n)^3}\left((n+1)\rho_{\mathrm{c}}^n-(n-1)\rho_{\mathrm{c}}^{n+1}-(n+1)\rho_{\mathrm{c}}+n-1\right)$$

Let $\chi(\rho_{\mathrm{c}})=(n+1)\rho_{\mathrm{c}}^n-(n-1)\rho_{\mathrm{c}}^{n+1}-(n+1)\rho_{\mathrm{c}}+n-1$. Then, $\chi'(\rho_{\mathrm{c}})=(n+1)\left(n\rho_{\mathrm{c}}^{n-1}+(1-n)\rho_{\mathrm{c}}^n-1\right)$ and $\chi''(\rho_{\mathrm{c}})=n(n^2-1)(1-\rho_{\mathrm{c}})\rho_{\mathrm{c}}^{n-2}$. Hence,

$$\left\{\begin{array}{ll}\chi''(\rho_{\mathrm{c}})>0, & \text{if } 0<\rho_{\mathrm{c}}<1\\ \chi''(\rho_{\mathrm{c}})<0, & \text{if } \rho_{\mathrm{c}}>1\end{array}\right.\Longrightarrow\left\{\begin{array}{ll}\chi'(\rho_{\mathrm{c}})<\chi'(1)=0, & \text{if } 0<\rho_{\mathrm{c}}<1\\ \chi'(\rho_{\mathrm{c}})<\chi'(1)=0, & \text{if } \rho_{\mathrm{c}}>1\end{array}\right.\Longrightarrow\left\{\begin{array}{ll}\chi(\rho_{\mathrm{c}})>\chi(1)=0, & \text{if } 0<\rho_{\mathrm{c}}<1\\ \chi(\rho_{\mathrm{c}})<\chi(1)=0, & \text{if } \rho_{\mathrm{c}}>1\end{array}\right..$$

Thus, $\left(\rho_{\mathrm{c}}^n\dfrac{n-n\rho_{\mathrm{c}}+\rho_{\mathrm{c}}^n-1}{(1-\rho_{\mathrm{c}}^n)^2}\right)'>0$ for $\rho_{\mathrm{c}}>0$ but $\rho_{\mathrm{c}}\neq1$. Moreover, by L'Hôpital's rule, $\lim\limits_{\rho_{\mathrm{c}}\to1}\left(\rho_{\mathrm{c}}^n\dfrac{n-n\rho_{\mathrm{c}}+\rho_{\mathrm{c}}^n-1}{(1-\rho_{\mathrm{c}}^n)^2}\right)'=(n^2-1)/(6n)\geq0$ with equality only if $n=1$ and $\rho_{\mathrm{c}}=1$. Consequently, $\rho_{\mathrm{c}}^n\dfrac{n-n\rho_{\mathrm{c}}+\rho_{\mathrm{c}}^n-1}{(1-\rho_{\mathrm{c}}^n)^2}$ is strictly increasing in ρ_{c}, which implies $\sum_{i=0}^{n-1}p_i(q^*)$ is strictly decreasing in γ.

(ii) When $q^*=1$, $\sum\limits_{i=0}^{n-1}p_i(q^*=1)=p_0(q^*=1)\sum\limits_{i=0}^{n-1}\rho^i=\left(\frac{1-\rho^n}{1-\rho}+\frac{\rho^n}{1-\rho+\gamma\rho}\right)^{-1}\frac{1-\rho^n}{1-\rho}$, which clearly is strictly increasing in γ. ■

Lemma 3.19 *For any $m=1,2,\ldots,n-1$, $\dfrac{z+z^2+\cdots+z^m}{1+z+\cdots+z^n}$ is strictly decreasing in $z\geq1$.*

Proof of Lemma 3.19. Let $p_i(z)=1+z+z^2+\cdots+z^i$, where $i=1,2,\ldots,n$.

Differentiate $\frac{p_m(z)-1}{p_n(z)}$ with respect to z,

$$\left(\frac{p_m(z)-1}{p_n(z)}\right)' = \left(\frac{z+z^2+\cdots+z^m}{1+z+\cdots+z^n}\right)' = \left(\frac{z^{m+1}-z}{z^{n+1}-1}\right)' = \frac{(m-n)z^{n+m+1}-(m+1)z^m+nz^{n+1}+1}{(z^{n+1}-1)^2}. \tag{3.11}$$

When $z=1$, by L'Hôpital's rule,

$$\left(\frac{p_m(z)-1}{p_n(z)}\right)'\Big|_{z=1} = \lim_{z\to 1}\frac{(m-n)z^{n+m+1}-(m+1)z^m+nz^{n+1}+1}{(z^{n+1}-1)^2} = \frac{m(m+1-n)}{2(n+1)} \le 0, \tag{3.12}$$

where the inequality is strict except $m=n-1$.

Consider $z>1$. Denote the numerator of (3.11) as $\psi_1(z) = (m-n)z^{n+m+1}-(m+1)z^m+nz^{n+1}+1$. Then, $\psi_1(z=1)=0$ and

$$\psi_1'(z) = z^{m-1}\left[(m-n)(n+m+1)z^{n+1}-(m+1)m+n(n+1)z^{n-m+1}\right]$$

with $\psi_1'(z=1)=0$. Moreover, let $\psi_2(z) = (m-n)(n+m+1)z^{n+1}-(m+1)m+n(n+1)z^{n-m+1}$. Then,

$$\psi_2'(z) = z^{n-m}\left[(m-n)(n+m+1)(n+1)z^m+n(n+1)(n-m+1)\right]$$

and $\psi_2'(z=1) = m(n+1)(m+1-n) \le 0$ since $m \le n-1$. It can be further shown that

$$\psi_2'(z)<0 \Leftrightarrow z > \left(\frac{n(n+1-m)}{(n-m)(n+1+m)}\right)^{1/m} \text{ and } \frac{n(n+1-m)}{(n-m)(n+1+m)} \le 1$$

since $m \le n-1$. Therefore, $\psi_2'(z)<0$ for $z>1$. Consequently, for $z>1$, we have

$$\left.\begin{matrix}\psi_2'(z)<0\\ \psi_2(z=1)=0\end{matrix}\right\} \Rightarrow \psi_2(z)<0 \Rightarrow \left.\begin{matrix}\psi_1'(z)<0\\ \psi_1(z=1)=0\end{matrix}\right\} \Rightarrow \psi_1(z)<0,$$

which indicates

$$\left(\frac{p_m(z)-1}{p_n(z)}\right)' < 0. \tag{3.13}$$

Combining (3.12) and (3.13), the derivative of $\frac{p_m(z)-1}{p_n(z)} \le 0$ with equality only when $m=n-1$ and $z=1$. Since the derivative being equal to 0 at one point does not affect the strict monotonicity of a function, we conclude that $\frac{z+z^2+\cdots+z^m}{1+z+\cdots+z^n}$ is strictly decreasing

in $z \geq 1$. ■

Lemma 3.20 *Let $y^*(\nu)$ be defined as in Corollary 3.3. Then, $\dfrac{z - \langle \nu \rangle}{1 + z + \cdots + z^n}$ decreases in $z \geq y^*(\nu)$.*

Proof of Lemma 3.20. To proceed, we need to first justify a structural property of $y^*(\nu)$. ■

Lemma 3.21 *If $\overline{\nu} = \nu + i$ for some $i \in \mathbb{N}$, $y^*(\nu) < y^*(\overline{\nu})$.*

Proof of Lemma 3.21. First, consider $\nu \geq 2$. For convenience, we write $y^*(\overline{\nu})$ and $y^*(\nu)$ as $y^*_{\overline{\nu}}$ and y^*_{ν} respectively. By the definition of $y^*(\nu)$ in Corollary 3.3,

$$n+1+\frac{1}{1-y^*\nu}-\frac{n+1}{1-(y^*\nu)^{n+1}}=\nu \text{ and } n+1+i+\frac{1}{1-y^*_{\overline{\nu}}}-\frac{n+1+i}{1-(y^*_{\overline{\nu}})^{n+1+i}}=\nu+i.$$

From the above two equations,

$$\frac{1}{1-y^*_{\overline{\nu}}}-\frac{n+1+i}{1-(y^*_{\overline{\nu}})^{n+1+i}}=\frac{1}{1-y^*\nu}-\frac{n+1}{1-(y^*\nu)^{n+1}}=\nu-(n+1). \tag{3.14}$$

We now consider the monotonicity of $\dfrac{1}{1-y}-\dfrac{k}{1-y^k}$ in k. Take the derivative in k,

$$\frac{\partial}{\partial k}\left(\frac{1}{1-y}-\frac{k}{1-y^k}\right)=\frac{y^k(1-k\ln(y))-1}{(1-y^k)^2}. \tag{3.15}$$

Recall that $\lim_{y\to 1} f(y) = n/2 + 1 \leq \nu$ for $\nu \geq 2$. Since $f(x)$ increases in y and $y^*(\nu)$ is the root to $f(y) = \nu$, we must have $y^*(\nu) \geq 1$ for $\nu \geq 2$. Thus, it is sufficient to focus our attention to the case $y^*(\nu) \geq 1$.

For $y = 1$,

$$\frac{\partial}{\partial k}\left(\frac{1}{1-y}-\frac{k}{1-y^k}\right)\bigg|_{y=1}=\lim_{y\to 1}\frac{y^k(1-k\ln(y))-1}{(1-y^k)^2}=-\frac{1}{2}<0.$$

For $y > 1$, since $\frac{\partial}{\partial y}\left(y^k(1-k\ln(y))-1\right) = -y^{k-1}k^2\ln(y) < 0$, then

$$\frac{\partial}{\partial k}\left(\frac{1}{1-y}-\frac{k}{1-y^k}\right)=\frac{y^k(1-k\ln(y))-1}{(1-y^k)^2}<\frac{1^k(1-k\ln(1))-1}{(1-y^k)^2}<0.$$

Therefore, $\frac{1}{1-y} - \frac{k}{1-y^k}$ is strictly decreasing in k for $y \geq 1$. It follows that

$$\frac{1}{1-y_\nu^*} - \frac{n+1+i}{1-(y_\nu^*)^{n+1+i}} < \frac{1}{1-y^*\nu} - \frac{n+1}{1-(y^*\nu)^{n+1}}. \tag{3.16}$$

The strict monotonicity of $f(y) = n+1+\frac{1}{1-y} - \frac{n+1}{1-y^{n+1}}$ implies that $\frac{1}{1-y} - \frac{n+1+i}{1-y^{n+1+i}}$ is strictly increasing in $y \geq 0$. Thus, by (3.14) and (3.16), it must be that $y^*\nu < y_\nu^*$.

Further, we establish that for $1 \leq \nu < 2$, $y_\nu^* < y_{\nu+1}^*$. Then by the above argument, we can conclude that $y_\nu^* < y_{\nu+1}^* \leq y_{\nu+i}^*$ for $i \in \mathbb{N}$. When $1 \leq \nu < 2$, we can obtain close-form formulae of $y_\nu^* = \frac{\nu-1}{2-\nu} = \frac{\langle\nu\rangle}{1-\langle\nu\rangle}$ and $y_{\nu+1}^* = \frac{(\nu+1)-2+\sqrt{4-3((\nu+1)-2)^2}}{2(3-(\nu+1))} = \frac{\langle\nu+1\rangle+\sqrt{4-3\langle\nu+1\rangle^2}}{2(1-\langle\nu+1\rangle)}$. Note that $0 \leq \langle\nu\rangle = \langle\nu+1\rangle < 1$. Under this condition, it can be easily verified that $y_\nu^* < y_{\nu+1}^*$. ■

Now, we are ready to show Lemma 3.20. Let $r \in (0,1)$ and write ν such that $\lfloor\nu\rfloor = j$ and $\langle\nu\rangle = r$, as ν_j^r. We will verify the monotonicity of $\frac{z-\langle\nu\rangle}{1+z+\cdots+z^n}$ for $z \geq y^*(\nu)$ in two steps: first, we show for $\nu_2^r \in (2,3)$, i.e., $n=2$, the term $\frac{z-\langle\nu_2^r\rangle}{1+z+z^2}$ is strictly decreasing in $z \geq y^*(\nu_2^r)$; second, we show for any given $n \geq 3$ and $\langle\nu_n^r\rangle = \langle\nu_2^r\rangle$, $\frac{z-\langle\nu_n^r\rangle}{1+z+\cdots+z^n}$ is strictly decreasing in $z \geq y^*(\nu_2^r)$ as well. Then, by Lemma 3.21, $y^*(\nu_n^r) > y^*(\nu_2^r)$. As a result, $\frac{z-\langle\nu_n^r\rangle}{1+z+\cdots+z^n}$ is strictly increasing in $z \geq y^*(\nu_n^r)$. This concludes (3.23) is strictly decreasing in $z \geq y^*(\nu_n^r)$ when $q^* = 0$ and hence $S_{\text{I}}(q^*=0)$ is strictly decreasing in γ.

Step 1. When $\nu_2^r \in (2,3)$, we can analytically solve $y^*(\nu_2^r)$ defined in Corollary 3.3 as $y^*(\nu_2^r) = \frac{\langle\nu_2^r\rangle+\sqrt{4-3\langle\nu_2^r\rangle^2}}{2(1-\langle\nu_2^r\rangle)}$. Thus, we only need to show $\frac{z-\langle\nu_2^r\rangle}{1+z+z^2}$ is strictly decreasing in $z \geq y^*(\nu_2^r)$. Differentiating $\frac{z-\langle\nu_2^r\rangle}{1+z+z^2}$ with respect to z, we have

$$\left(\frac{z-\langle\nu_2^r\rangle}{1+z+z^2}\right)' = \frac{-z^2+2\langle\nu_2^r\rangle z+\langle\nu_2^r\rangle+1}{(1+z+z^2)^2} < 0 \Leftrightarrow z > \langle\nu_2^r\rangle + \sqrt{\langle\nu_2^r\rangle^2+\langle\nu_2^r\rangle+1}.$$

We observe that $\frac{\langle\nu_2^r\rangle+\sqrt{4-3\langle\nu_2^r\rangle^2}}{2(1-\langle\nu_2^r\rangle)} > \langle\nu_2^r\rangle+\sqrt{\langle\nu_2^r\rangle^2+\langle\nu_2^r\rangle+1}$ for $\langle\nu_2^r\rangle \in (0,1)$. Hence, $\frac{z-\langle\nu_2^r\rangle}{1+z+z^2}$ is strictly decreasing in $z \geq y^*(\nu_2^r) = \frac{\langle\nu_2^r\rangle+\sqrt{4-3\langle\nu_2^r\rangle^2}}{2(1-\langle\nu_2^r\rangle)}$ and so is $S_{\text{I}}(q^*=0)$ in γ.

Step 2. Since $\frac{z-\langle\nu_2^r\rangle}{1+z+z^2}$ is strictly decreasing for $z \geq y^*(\nu_2^r)$, at a given $\overline{z} \geq y^*(\nu_2^r)$ we have

$$\left(\frac{z-\langle\nu_2^r\rangle}{1+z+z^2}\right)'\Big|_{z=\overline{z}} = \left(\frac{z-\langle\nu_2^r\rangle}{p_2(z)}\right)'\Big|_{z=\overline{z}} = \frac{p_2(\overline{z})-(\overline{z}-\langle\nu_2^r\rangle)p_2'(\overline{z})}{p_2^2(\overline{z})} < 0 \Leftrightarrow \langle\nu_2^r\rangle < \overline{z} - \frac{p_2(\overline{z})}{p_2'(\overline{z})}.$$

On the other hand, because $\langle \nu_n^r \rangle = \langle \nu_2^r \rangle$, thus at the same $\overline{z}$,

$$\begin{aligned}\left(\frac{z-\langle \nu_n^r\rangle}{1+z+\cdots+z^n}\right)'\Big|_{z=\overline{z}} = \left(\frac{z-\langle \nu_n^r\rangle}{p_n(z)}\right)'\Big|_{z=\overline{z}} &= \frac{p_n(\overline{z})-(\overline{z}-\langle \nu_n^r\rangle)p_n'(\overline{z})}{p_n^2(\overline{z})} \\ &< \frac{p_n(\overline{z})-\left(\overline{z}-\left(\frac{p_2(\overline{z})}{p_2'(\overline{z})}\right)\right)p_n'(\overline{z})}{p_n^2(\overline{z})} \\ &= \frac{p_2'(\overline{z})p_n(\overline{z})-p_2(\overline{z})p_n'(\overline{z})}{p_n^2(\overline{z})p_2'(\overline{z})} = \frac{1}{p_2'(\overline{z})}\left(\frac{p_2(\overline{z})}{p_n(\overline{z})}\right)' < 0,\end{aligned}$$

where the last inequality is because $p_2'(\overline{z}) > 0$ and

$$\frac{p_2(z)}{p_n(z)} = \frac{1+z+z^2}{1+z+\cdots+z^n} = \frac{1}{1+z+\cdots+z^n} + \frac{z+z^2}{1+z+\cdots+z^n}$$

is strictly decreasing in $z \geq 1$ when $n \geq 3$. Since $y^*(\nu_2^r) \geq 1$, then $\frac{z-\langle \nu_n^r\rangle}{1+z+\cdots+z^n}$ is strictly decreasing in $z \geq y^*(\nu_2^r)$. ■

3.8.2 Proofs

Proof of Lemma 3.1. We first demonstrate a structural property of $p_i(q)$: There exists q-dependent $k \in \mathbb{N}$ such that $p_i(q)$ is decreasing in q for all $0 \leq i < k$ and $p_i(q)$ is strictly increasing in q for all $i \geq k$. For $\gamma \in [0,1)$, $\lambda_U > 0$. By (3.5), it is clear that $p_0(q)$ is strictly decreasing in q. Moreover, since $\sum_{i=0}^{\infty} p_i(q) = 1$, there must exist some $i' \in \mathbb{N}$ such that $p_{i'}(q)$ is strictly increasing in q. Let $k = \min\{i \in \mathbb{N} \mid p_i(q) \text{ is strictly increasing in } q\}$. By the balance equations (3.1) and (3.2), since $p_k(q)$ is strictly increasing in q, $p_{k+1}(q)$ is strictly increasing in q, and recursively, $p_i(q)$ is strictly increasing in q for all $i \geq k$. By the definition of $p_k(q)$, all $p_i(q)$'s for $0 \leq i < k$ decrease in q.

Now we use the property above to show the stochastic monotonicity of Q. Fix $l \in \mathbb{N}$. If $l \leq k$, $\mathbb{P}(Q(q) \geq l) = 1 - \mathbb{P}(Q(q) < l) = 1 - (\sum_{i=0}^{l-1} p_i(q))$ increases in q, because $p_i(q)$, for $i \leq l-1 \leq k-1$, decreases in q by part (i). If $l > k$, $\mathbb{P}(Q(q) \geq l) = \sum_{i=l}^{\infty} p_i(q)$ strictly increases in q, because $p_i(q)$, for $i \geq l > k$, strictly increases in q. The result follows by the definition of the usual stochastic order in Shaked and Shanthikumar (2007). Thus, for any $0 \leq q_1 < q_2 \leq 1$, $Q(q_1) \leq_{\text{st}} Q(q_2)$, which implies that $\mathbb{E}(Q(q_1)) \leq \mathbb{E}(Q(q_2))$. By Theorem 1.A.8. in Shaked and Shanthikumar (2007), the inequality must be strict, i.e., $\mathbb{E}(Q(q_1)) < \mathbb{E}(Q(q_2))$ for $q_1 < q_2$. Since $W(q) = \mathbb{E}(Q(q))/\mu$, $W(q_1) < W(q_2)$. ■

Proof of Proposition 3.2. The proof follows from similar arguments to Hassin and Haviv (2003) p. 46. ■

Proof of Corollary 3.3. By Proposition 3.2(i), no uninformed customers join the queue if and only if $cW(0) \geq R$. Plugging in (3.6), we have $cW(0) \geq R \iff f(\gamma\rho) \geq \nu$, where

$f(y) \equiv n+1+\dfrac{1}{1-y}-\dfrac{n+1}{1-y^{n+1}}, y \geq 0$. The result follows if (1) $f(y)$ is continuous and strictly increasing and (2) $f(y)=\nu$ has a unique solution. Since $\lim_{y\to 1} f(y) = n/2+1$, the continuity is guaranteed. The monotonicity of $f(y)$ results from the fact that

$$f'(y) = \frac{1}{(1-y)^2} - \frac{(n+1)^2 y^n}{(1-y^{n+1})^2} \geq \frac{1}{(1-y)^2} - \frac{(n+1)^2}{(1-y^{n+1})^2}\left(\frac{1}{n+1}\cdot\frac{1-y^{n+1}}{1-y}\right)^2 = 0,$$

where the last inequality is due to the inequality of arithmetic and geometric means

$$y^{n/2} = \sqrt[n+1]{1\times y\times y^2\times\cdots\times y^n} \leq \frac{1+y+y^2+\cdots+y^n}{n+1} = \frac{1}{n+1}\cdot\frac{1-y^{n+1}}{1-y}.$$

Note that $f'(y)=0$ only at a *single* point $y=1$. Thus, $f(y)$ is strictly increasing in y. At last, since $f(0)=1$, $\lim_{y\to\infty} f(y) = n+1$ and $\nu \in [n, n+1)$, $n \geq 1$. Therefore, $f(y)=\nu$ has a unique solution $y^*(\nu) \geq 0$. ■

Proof of Corollary 3.4. By Proposition 3.2(ii) and (3.6), uninformed customers all join the queue if and only if

$$cW(1) \leq R \Longleftrightarrow \left(\frac{1}{1-(1-\gamma)\rho}\right)^2 - \langle\nu\rangle\left(\frac{1}{1-(1-\gamma)\rho}\right) - L(\rho,\nu) \leq 0,$$

where

$$L(\rho,\nu) \equiv \frac{\langle\nu\rangle\sum_{i=1}^{n-1}\rho^i + \sum_{i=1}^{n-1}\sum_{j=0}^{i-1}\rho^j}{\rho^n} \geq 0.$$

Note that by (3.3),

$$0 \leq \mathbb{P}(Q\geq n) = \sum_{i=n}^{\infty} p_i(q) = \sum_{i=n}^{\infty}\left(\rho_{\mathfrak{c}}(q)\right)^n\left(\rho_{\mathfrak{v}}(q)\right)^{i-n} p_0(q) = \left(\rho_{\mathfrak{c}}(q)\right)^n\left(1-\rho_{\mathfrak{v}}(q)\right)^{-1} p_0(q). \tag{3.17}$$

Then $(1-\rho_{\mathfrak{v}}(q))^{-1} \geq 0$ for all q. In particular, if $q=1$, $(1-\rho_{\mathfrak{v}}(q=1))^{-1} = \frac{1}{1-(1-\gamma)\rho} \geq 0$. Therefore,

$$\begin{aligned} cW(1) \leq R &\iff \left(\frac{1}{1-(1-\gamma)\rho}\right)^2 - \langle\nu\rangle\left(\frac{1}{1-(1-\gamma)\rho}\right) - L(\rho,\nu) \leq 0 \\ &\iff 0 \leq \frac{1}{1-(1-\gamma)\,\rho} \leq \frac{\langle\nu\rangle + \sqrt{\langle\nu\rangle^2 + 4L(\rho,\nu)}}{2}. \end{aligned}$$

The rest of the proof follows by solving for the condition on γ from the above inequality. ■

Proof of Theorem 3.5. We first demonstrate how the two critical information levels

$\gamma_0^*(\rho,\nu)$ and $\gamma_1^*(\rho,\nu)$ change w.r.t. the offered load ρ. It is obvious that $\gamma_0^*(\rho,\nu) = y^*(\nu)/\rho$ is strictly decreasing in ρ. Since $\lim_{\rho\to\infty}\gamma_0^*(\rho,\nu) = 0$ and $\lim_{\rho\to y^*(\nu)}\gamma_0^*(\rho,\nu) = 1$, we claim $0 \le \gamma_0^*(\rho,\nu) < 1$ if and only if $\rho > y^*(\nu)$; and $\gamma_0^*(\rho,\nu) \ge 1$ if and only if $\rho \le y^*(\nu)$. On the other hand, it is easy to show that $\gamma_1^*(\rho,\nu) \ge 0 \Leftrightarrow \rho \ge 1 - 1/\nu \ge 0$ because $L(\rho,\nu) \ge 0$ and $\nu \ge 1$. Due to the same fact that $L(\rho,\nu) \ge 0$,

$$\gamma_1^*(\rho,\nu) \le 1 \iff (\langle\nu\rangle(\rho-1)+2-\rho)(\rho^{n+1}-1) \ge (n+1)(\rho-1). \tag{3.18}$$

The inequality always holds if $\rho = 1$. We then only discuss the case $\rho \ne 1$. Dividing both sides by $(1-\rho)(1-\rho^{n+1}) > 0$, (3.18) can be equivalently transformed as

$$\gamma_1^*(\rho,\nu) \le 1 \Longleftrightarrow \frac{\langle\nu\rangle(\rho-1)+1+1-\rho}{1-\rho} \le \frac{n+1}{1-\rho^{n+1}} \Longleftrightarrow n+1+\frac{1}{1-\rho}-\frac{n+1}{1-\rho^{n+1}} \le \nu \Longleftrightarrow \rho \le y^*(\nu),$$

where the last equivalence results from the fact $f(y) = n+1+\frac{1}{1-y}-\frac{n+1}{1-y^{n+1}}$ is increasing in y.

(i) Consider $0 \le \rho < 1 - 1/\nu$. In this case, $\gamma_1^*(\rho,\nu) < 0$. By Corollary 3.4, $q^* = 1$ for all $\gamma \in [0,1]$.

(ii) Consider $1 - 1/\nu \le \rho \le y^*(\nu)$. In this case, $\gamma_0^*(\rho,\nu) = y^*(\nu)/\rho \ge 1$. Thus, $q^* \ne 0$ for all $\gamma \in [0,1]$ by Corollary 3.3. However, $\gamma_1^*(\rho,\nu) \in [0,1]$ for $1 - 1/\nu \le \rho \le y^*(\nu)$, which implies that $q^* = 1$ for $\gamma_1^*(\rho,\nu) \le \gamma \le 1$. By the uniqueness of q^*, $0 < q^* < 1$ for $0 \le \gamma < \gamma_0^*(\rho,\nu)$.

(iii) Consider $\rho > y^*(\nu)$. In this case, $0 \le \gamma_0^*(\rho,\nu) < 1$. Therefore, $q^* = 0$ for $\gamma_0^*(\rho,\nu) \le \gamma \le 1$. Then, for $0 \le \gamma < \gamma_0^*(\rho,\nu)$, it is only possible that $0 < q^* \le 1$. However, by the uniqueness of q^*, it can be easily shown, by contradiction, that it must be that $0 < q^* < 1$ for $0 \le \gamma < \gamma_0^*(\rho,\nu)$. ■

Proof of Lemma 3.6. We consider two cases separately: (i) $q^*(\gamma) = 0$ and (ii) $q^*(\gamma) \in (0,1)$.

(i) $q^*(\gamma) = 0$. By (3.5), $p_0(\gamma) = \left(1 + \sum_{i=0}^{n-1}(\gamma\rho)^{i+1}\right)^{-1}$, which is strictly decreasing in γ.

(ii) $q^*(\gamma) \in (0,1)$. Since it is difficult to analytically solve q^* as a function of γ from $cW(q) = R$, we have to verify the result indirectly. Specifically, we focus on $\rho_c = \rho[\gamma + q^*(1-\gamma)]$ instead of q^* for the monotonicity of $p_0(\gamma)$. Since we are only interested in the nontrivial cases where $\rho > 0$ and $\gamma > 0$, $\rho_c > 0$. It is sufficient to show that $\frac{d\rho_c}{d\gamma} > 0$ and $\frac{dp_0}{d\rho_c} < 0$. Then we can obtain $\frac{dp_0}{d\gamma} = \frac{dp_0}{d\rho_c}\frac{d\rho_c}{d\gamma} < 0$ for $q^* \in (0,1)$.

First, we verify $d\rho_c/d\gamma > 0$. When $q^* \in (0,1)$, the relationship between ρ_c and γ is

determined by the equilibrium equation $cW(q^*) = \frac{c}{\mu}\sum_{i=0}^{\infty}(i+1)p_i(q^*) = R$. By (3.6),

$$\begin{aligned} cW(q^*) &= \frac{c}{\mu}\sum_{i=0}^{\infty}(i+1)p_i(q^*) \\ &= \frac{c}{\mu}p_0(q^*)\left[\frac{1-\rho_c^n}{1-\rho_c} + \frac{\rho_c}{(1-\rho_c)^2} + \rho_c^n\left(\frac{1-n}{1-\rho_c} - \frac{1}{(1-\rho_c)^2} + \frac{1}{(1-\rho_c+\gamma\rho)^2} + \frac{n}{1-\rho_c+\gamma\rho}\right)\right] \\ &= R, \end{aligned} \tag{3.19}$$

where $p_0(q^*) = \left(\frac{1-\rho_c^n}{1-\rho_c} + \frac{\rho_c^n}{1-\rho_c+\gamma\rho}\right)^{-1}$. Define $\langle\nu\rangle = \nu - n$. Then Eq. (3.19) is equivalent to

$$\frac{\langle\nu\rangle(\rho_c-1)\rho_c^n + \nu - \nu\rho_c + \rho_c^n - 1}{(1-\rho_c)^2\rho_c^n} = \frac{1-\langle\nu\rangle(1-\rho_c+\gamma\rho)}{(1-\rho_c+\gamma\rho)^2}, \tag{3.20}$$

where the right hand side is exactly $L(\rho_c)$ defined in Corollary 3.4. Recall from Eq.(3.4) that $1-\rho_u(q^*) = 1-q^*\lambda_u/\mu = 1-\rho_c+\gamma\rho \geq 0$. Thus, Eq.(3.20) gives rise to the only positive solution to $(1-\rho_c+\gamma\rho)$, which further leads to a unique expression of $\gamma(\rho_c)$. In other words, we know from Eq.(3.20) that

$$1-\rho_c+\gamma\rho = \frac{-\langle\nu\rangle + \sqrt{\langle\nu\rangle^2 + 4L(\rho_c)}}{2L(\rho_c)} \Longleftrightarrow \gamma(\rho_c) = \frac{1}{\rho}\left(2\left(\langle\nu\rangle + \sqrt{\langle\nu\rangle^2+4L(\rho_c)}\right)^{-1} + \rho_c - 1\right). \tag{3.21}$$

It can be shown that $L(\rho_c)$ is strictly decreasing in ρ_c, i.e., $dL/d\rho_c < 0$ (see Lemma 3.16 in the Online Appendix B). Therefore, $\gamma(\rho_c)$ in (3.21) is strictly increasing in ρ_c, i.e., $d\gamma/d\rho_c > 0$. Moreover, since the inverse function of a strictly increasing function is also strictly increasing, $d\rho_c/d\gamma > 0$.

Second, we show $dp_0/d\rho_c < 0$. We write $p_0(q^*)$ as a function of ρ_c:

$$p_0(q^*) \overset{(3.5)}{=} \left(\frac{1-\rho_c^n}{1-\rho_c} + \frac{\rho_c^n}{1-\rho_c+\gamma\rho}\right)^{-1} \overset{(3.21)}{=} \left[\sum_{i=0}^{n-1}\rho_c^i + \frac{\rho_c^n}{2}\langle\nu\rangle + \sqrt{\rho_c^{2n}\langle\nu\rangle^2/4 + \rho_c^{2n}L(\rho_c)}\right]^{-1}.$$

Recall that $\rho_c > 0$. Hence, if $\rho_c^{2n}L(\rho_c)$ is strictly increasing in ρ_c, we can easily show that $p_0(q^*)$ is strictly decreasing in ρ_c. Note

$$\rho_c^{2n}L(\rho_c) = \rho_c^n\frac{\langle\nu\rangle(1-\rho_c)(1-\rho_c^n) + (1-\rho_c)\sum_{i=0}^{n-1}(1-\rho_c^i)}{(1-\rho_c)^2} = \rho_c^n\left(\langle\nu\rangle\sum_{i=0}^{n-1}\rho_c^i + \sum_{i=1}^{n-1}\sum_{j=0}^{i-1}\rho_c^j\right), \tag{3.22}$$

which confirms that $\rho_c^{2n}L(\rho_c)$ is indeed strictly increasing in $\rho_c > 0$ and implies that $dp_0/d\rho_c < 0$. ■

Proof of Theorem 3.7. (i) Recall that $\lambda(q) = \mu(1-p_0(q))$. Hence, for $q^* \in [0,1)$, it is obvious that $\lambda(q^*)$ strictly increases in γ, since we have shown that $p_0(q^*)$ strictly decreases in γ in Lemma 3.6.

(ii) If $q^*=1$, $\rho_{\text{c}}(q^*=1)=\rho$ and $\rho_{\text{u}}(q^*=1)=(1-\gamma)\rho$. Then

$$\begin{aligned}\lambda(q^*=1) &\overset{(3.7)}{=} \left(\sum_{i=0}^{n-1} p_i(1)\right)\lambda_{\text{I}}+\lambda_{\text{u}} \\ &\overset{(3.3),(3.5)}{=} \left(\frac{1-\rho^n}{1-\rho}+\frac{\rho^n}{1-(1-\gamma)\rho}\right)^{-1}\left(\frac{1-\rho^n}{1-\rho}\right)\gamma\Lambda+(1-\gamma)\Lambda \\ &= \frac{-1+\rho-\gamma\rho+\gamma\rho^n}{-1+\rho-\gamma\rho+\gamma\rho^{n+1}}\Lambda.\end{aligned}$$

Differentiating $\lambda(q^*=1)$ w.r.t. γ, we have $\frac{d}{d\gamma}\lambda(q^*=1)=-\frac{\rho^n(1-\rho)^2}{(-1+\rho-\gamma\rho+\gamma\rho^{n+1})^2}\Lambda<0$ for $\rho>0$. Thus, $\lambda(q^*=1)$ is strictly decreasing in γ when $q^*=1$. ■

Proof of Corollary 3.8. The result immediately follows by combining Theorems 3.5 and 3.7. ■

Proof of Theorem 3.9. (i) We first verify that $d\overline{S}_{\text{I}}(q^*(\gamma'))/d\gamma<0$ for any given γ' such that $0\le q^*(\gamma')<1$. Lemma 3.17 in the Online Appendix B states that for any such γ' there exists $k\le n-1$ such that $dp_i(q^*(\gamma'))/d\gamma<0$ for $0\le i\le k$ and $dp_i(q^*(\gamma'))/d\gamma\ge 0$ for $k<i<n$. Therefore,

$$\begin{aligned}\frac{d\overline{S}_{\text{I}}(q^*(\gamma'))}{d\gamma} &\le \sum_{i=0}^{k}\frac{dp_i(q^*(\gamma'))}{d\gamma}\left(R-c\frac{k+1}{\mu}\right)+\sum_{i=k+1}^{n-1}\frac{dp_i(q^*(\gamma'))}{d\gamma}\left(R-c\frac{k+1}{\mu}\right) \\ &= \left(R-c\frac{k+1}{\mu}\right)\sum_{i=0}^{n-1}\frac{dp_i(q^*(\gamma'))}{d\gamma}<0,\end{aligned}$$

where the last inequality results from Proposition 3.18(i) (see Online Appendix B).

We next show that $\overline{S}_{\text{I}}(q^*)$ strictly increases in γ for $q^*=1$. Assume that $\gamma_1\ge\gamma_2$,

$$\begin{aligned}\overline{S}_{\text{I}}(q^*(\gamma_2))-\overline{S}_{\text{I}}(q^*(\gamma_1)) &= \sum_{i=0}^{n-1}\left(p_i(q^*(\gamma_2))-p_i(q^*(\gamma_1))\right)\left(R-c\frac{i+1}{\mu}\right) \\ &\le \sum_{i=0}^{n-1}\left(p_i(q^*(\gamma_2))-p_i(q^*(\gamma_1))\right)\left(R-c\frac{1}{\mu}\right)<0,\end{aligned}$$

where the last inequality is due to Proposition 3.18(ii), i.e., $\sum_{i=0}^{n-1}p_i(q^*=1)$ strictly increases in γ.

(ii) By definition, $\overline{S}_{\text{u}}(q^*)=0$ if $0\le q^*<1$. Thus, we only need to consider the case where $q^*=1$,

$$\overline{S}_{\text{u}}(q^*=1) = \sum_{i=0}^{\infty}p_i(q^*=1)\left(R-c\frac{i+1}{\mu}\right)$$

$$= \quad p_0(q^*=1)\left(\sum_{i=0}^{n-1}\rho^i\left(R-c\frac{i+1}{\mu}\right)+\sum_{i=n}^{\infty}\rho^n((1-\gamma)\rho)^{i-n}\left(R-c\frac{i+1}{\mu}\right)\right).$$

First, $p_0(q^*=1)=\left(\frac{1-\rho^n}{1-\rho}+\frac{\rho^n}{1-\rho+\gamma\rho}\right)^{-1}$ strictly increases in γ. Second, $R-c\frac{i+1}{\mu}<0$ for $i\geq n$, which implies $\rho^n((1-\gamma)\rho)^{i-n}\left(R-c\frac{i+1}{\mu}\right)$ increases in γ. Therefore, $\overline{S}_{\mathrm{U}}(q^*=1)$ strictly increases in γ. ■

Proof of Theorem 3.10. We prove the results case by case for $q^*=0$, $q^*\in(0,1)$ and $q^*=1$, respectively. By definition, $S_{\mathrm{U}}(q^*)=0$ for all γ if $q^*=0$. By Proposition 3.2(iii), $q^*\in(0,1)$ must satisfy that $R=cW(q^*)$, hence, $S_{\mathrm{U}}(q^*)=0$ for all γ if $q^*\in(0,1)$. Then, $S_{\mathrm{U}}(q^*)>0$ can only happen when $q^*=1$.

(a) When $q^*=0$, we have $S_{\mathrm{I}}(q^*=0)=\left(\sum_{i=0}^{n-1}p_i(0)\left(R-c\frac{i+1}{\mu}\right)\right)\cdot\gamma\Lambda=(g(\gamma\rho)+\langle\nu\rangle)\cdot\mu^2/c$, where

$$g(z)\equiv\frac{1+nz-\langle\nu\rangle(1-z)}{1-z^{n+1}}-\frac{1}{1-z}.$$

If $n=1$, $S_{\mathrm{I}}(q^*=0)=\left(-\frac{\langle\nu\rangle}{1+\gamma\rho}+\langle\nu\rangle\right)\frac{\mu^2}{c}$, which strictly increases in γ. So does $S_{\mathrm{I}}+S_{\mathrm{U}}$ for $1\leq\nu<2$.

Next, we consider the case of $n\geq 2$. By Corollary 3.3, $q^*=0$ if and only if $z\equiv\gamma\rho\geq y^*(\nu)$. Moreover, it can be shown that $y^*(\nu)\geq 1$ for $\nu\geq 2$: By the definition of $y^*(\nu)$, we have $f(y)=\nu$; note that $\lim_{y\to 1}f(y)=n/2+1\leq\nu$ for $n\geq 2$; since $f(y)$ increases in y, $y^*(\nu)\geq 1$ for $\nu\geq 2$. Thus, we only need to show the monotonicity of $S_{\mathrm{I}}(q^*=0)=(g(z)+\langle\nu\rangle)\cdot\mu^2/c$, or equivalently $g(z)$, in z in the domain of $z\geq y^*(\nu)\geq 1$. We rewrite:

$$g(z)=\frac{\sum_{m=2}^{n-1}(z+z^2+\cdots+z^m)}{1+z+\cdots+z^n}+\frac{z-\langle\nu\rangle}{1+z+\cdots+z^n}. \tag{3.23}$$

It can be shown that $\frac{z+z^2+\cdots+z^m}{1+z+\cdots+z^n}$, $m=1,2,\ldots,n-1$, strictly decreases in $z\geq 1$ (see Lemma 3.19 in the Online Appendix B). Then,

• If $\langle\nu\rangle=0$, every term in (3.23) strictly decreases in $z\geq 1$, and hence so does $S_{\mathrm{I}}(q^*=0)$ in γ.

• If $\langle\nu\rangle\in(0,1)$, all terms but the last one in (3.23) are strictly decreasing in $z\geq 1$. Nevertheless, we can show that $\frac{z-\langle\nu\rangle}{1+z+\cdots+z^n}$ decreases in $z\geq y^*(\nu)$. Therefore, $S_{\mathrm{I}}(q^*=0)$ strictly decreases in γ.

(b) When $q^*\in(0,1)$, again let $\rho_{\mathrm{c}}\equiv\rho(\gamma+q^*(1-\gamma))$. Then, we have

$$S_{\mathrm{I}}(q^*)=\overline{S}_{\mathrm{I}}(q)\cdot\gamma\Lambda=\frac{c\Lambda}{\mu}p_0(q^*)\left(\nu\frac{1-\rho_{\mathrm{c}}^n}{1-\rho_{\mathrm{c}}}-\frac{1-(n+1)\rho_{\mathrm{c}}^n+n\rho_{\mathrm{c}}^{n+1}}{(1-\rho_{\mathrm{c}})^2}\right)\gamma(\rho_{\mathrm{c}})\overset{(3.20)}{=}c\rho p_0(q^*)\rho_{\mathrm{c}}^nL(\rho_{\mathrm{c}})\gamma(\rho_{\mathrm{c}}).$$

It can be further shown[10] that

$$S_{\text{I}}(q^*) = c\rho p_0(q^*)\rho_{\text{c}}^n L(\rho_{\text{c}})\gamma(\rho_{\text{c}}) = 1 - \nu p_0(q^*).$$

By Lemma 3.6, $p_0(q^*)$ strictly decreases in γ when $q^* \in (0,1)$. Therefore, S_{I} is strictly increasing in γ.

(c) Lastly, when $q^* = 1$, $\rho_{\text{c}} = \rho$. Then, we have

$$\begin{aligned} S_{\text{I}}(q^*=1) + S_{\text{U}}(q^*=1) &= \left(\sum_{i=0}^{n-1} p_i(1)\left(R - c\tfrac{i+1}{\mu}\right)\right)\cdot\gamma\Lambda + \left(\sum_{i=0}^{\infty} p_i(1)\left(R - c\tfrac{i+1}{\mu}\right)\right)\cdot(1-\gamma)\Lambda \\ &= \frac{c\Lambda}{\mu}p_0(1)\left[\left(\frac{n(1-\rho)-(1-\rho^n)}{(1-\rho)^2} + \langle\nu\rangle\frac{1-\rho^n}{1-\rho}\right)\right. \\ &\qquad \left. +\rho^n\left(\frac{\gamma-1}{(1-\rho+\gamma\rho)^2} + \langle\nu\rangle\frac{1-\gamma}{1-\rho+\gamma\rho}\right)\right]. \end{aligned}$$

First, $p_0(q^*=1) = \left(\frac{1-\rho^n}{1-\rho} + \frac{\rho^n}{1-(1-\gamma)\rho}\right)^{-1}$ strictly increases in γ. Second, because $S_{\text{I}}(q^*=1) + S_{\text{U}}(q^*=1) > 0$, the term in the square bracket is positive. Moreover, it also strictly increases in γ, since

$$\frac{\partial}{\partial\gamma}\left(\frac{\gamma-1}{(1-\rho+\gamma\rho)^2} + \langle\nu\rangle\frac{1-\gamma}{1-\rho+\gamma\rho}\right) = \frac{1-\langle\nu\rangle+(1+\langle\nu\rangle)(1-\gamma)\rho}{(1-\rho+\gamma\rho)^3} \geq 0,$$

where the last inequality results from the fact that $1-\rho+\gamma\rho > 0$ implied by (3.17) and $\langle\nu\rangle \in [0,1)$. As a result, $S_{\text{I}}(q^*=1) + S_{\text{U}}(q^*=1)$ is strictly increasing in γ. ■

Proof of Corollary 3.11. The result immediately follows by combining Theorems 3.5 and 3.10. ■

Proof of Proposition 3.12. Customers evaluate their options, between being informed after paying a fee f and staying uninformed, and then pick the one that maximizes their net utility. Note that $\overline{S}_{\text{I}}(\gamma_1^*) - f = \overline{S}_{\text{U}}(\gamma_1^*) = 0$. Therefore, no one would have an incentive to deviate at $\gamma = \gamma_1^*$ and thus $\gamma = \gamma_1^*$ is an equilibrium. We will show that at any information level $\gamma \neq \gamma_1^*$, either the informed or the uninformed ones have an incentive to deviate.

Assume that $0 \leq \gamma < \gamma_1^*$. Recall that $\rho \in [\underline{\rho}, \overline{\rho}]$, $0 < q^*(\gamma) < 1$ and then $\overline{S}_{\text{U}}(\gamma_1^*) = 0$ for $0 < \gamma < \gamma_1^*$. By Theorem 3.9, $0 = \overline{S}_{\text{I}}(\gamma_1^*) - f < \overline{S}_{\text{I}}(\gamma) - f$, with the latter decreasing in γ. Thus, $\overline{S}_{\text{U}}(\gamma) < \overline{S}_{\text{I}}(\gamma) - f$ for $0 < \gamma < \gamma_1^*$. Then, uninformed customers would have an incentive to deviate and pay the information access fee f to become informed.

If $\gamma_1^* < \gamma \leq 1$, it is the informed individuals who want to deviate. To see this, we will show that $\overline{S}_{\text{I}}(\gamma) - f < \overline{S}_{\text{U}}(\gamma)$ for $\gamma_1^* < \gamma \leq 1$. Note that in this range of γ, $q^* = 1$. We

[10] The derivation can be considered as a special case of Eq. (3.32) in the proof of Theorem 3.15.

have

$$\begin{aligned}\overline{S}_{\text{I}}(\gamma) - f - \overline{S}_{\text{U}}(\gamma) &= \sum_{i=0}^{n-1} p_i\left(q^*=1\right)\left(R - c\frac{i+1}{\mu}\right) - f - \sum_{i=0}^{\infty} p_i\left(q^*=1\right)\left(R - c\frac{i+1}{\mu}\right) \\ &= \frac{c}{\mu}\cdot\frac{1-(v-n)\left(1-\rho+\gamma\rho\right)}{\left(1-\rho+\gamma\rho\right)\left(1+\gamma\rho\frac{(\rho^n-1)}{\rho-1}\right)}\rho^n - f.\end{aligned}$$

Since f is a constant, $\overline{S}_{\text{I}}(\gamma) - f - \overline{S}_{\text{U}}(\gamma)$ apparently decreases in γ. Thus, for $\gamma_1^* < \gamma \leq 1$, $\overline{S}_{\text{I}}(\gamma) - f - \overline{S}_{\text{U}}(\gamma) < \overline{S}_{\text{I}}(\gamma_1^*) - f - \overline{S}_{\text{U}}(\gamma_1^*) = 0 \iff \overline{S}_{\text{I}}(\gamma) - f < \overline{S}_{\text{U}}(\gamma)$. ■

Proof of Proposition 3.13. The proof follows the same idea as that of Proposition 3.12. We thus omit the details. ■

Proof of Theorem 3.14. First, following the same approach in the proof of Lemma 3.1, it can be shown that the expected sojourn time $W(q)$ in the heterogenous case strictly increases in q too. As a result, it is easy to further demonstrate that there exists a unique joining equilibrium $q^* \in [0,1]$ for uninformed customers. In the case of $q^* = 0$ or 1, the demonstration of the monotonicity of $\lambda(q^*)$ is parallel to that of the homogeneous reward case, in which $R_{\text{I}} = R_{\text{U}} = R$ and $c_{\text{I}} = c_{\text{U}} = c$. Thus, for the rest of the proof, we only consider the cases in which $q^* \in (0,1)$.

Since $\lambda(q) = \mu(1-p_0(q))$, the monotonicity of $\lambda(q^*)$ in γ is opposite to that of $p_0(q^*)$. Thus, instead of directly proving that $\lambda(q^*)$ is strictly increasing in γ, we will show that $p_0(q^*)$ is strictly decreasing in γ in two steps: (i) From the expression of $R_{\text{U}} = c_{\text{U}}W(q)$, derive γ as a function of ρ_{C} and prove $\frac{d\rho_{\text{C}}}{d\gamma} > 0$; (ii) From the expression of $p_0(q)$, prove $\frac{dp_0}{d\rho_{\text{C}}} < 0$. Then, combining these two results, we obtain $\frac{dp_0}{d\gamma} = \frac{dp_0}{d\rho_{\text{C}}}\frac{d\rho_{\text{C}}}{d\gamma} < 0$.

Step (i). Rewrite $R_{\text{U}} = c_{\text{U}}W(q)$ as

$$H\left(\rho_{\text{C}}\right)\left(1-\rho_{\text{C}}+\gamma\rho\right)^2 + \left(\nu_{\text{U}} - n_{\text{I}}\right)\left(1-\rho_{\text{C}}+\gamma\rho\right) - 1 = 0, \tag{3.24}$$

where

$$H\left(\rho_{\text{C}}\right) = \frac{\left(\nu_{\text{U}}-n_{\text{I}}\right)\left(\rho_{\text{C}}-1\right)\rho_{\text{C}}^{n_{\text{I}}} + \nu_{\text{U}} - \nu_{\text{U}}\rho_{\text{C}} + \rho_{\text{C}}^{n_{\text{I}}} - 1}{\left(\rho_{\text{C}}-1\right)^2\rho_{\text{C}}^{n_{\text{I}}}} = \frac{\left(\nu_{\text{U}}-n_{\text{I}}\right)\sum_{i=0}^{n_{\text{I}}-1}\rho_{\text{C}}^i}{\rho_{\text{C}}^{n_{\text{I}}}} + \frac{\sum_{i=1}^{n_{\text{I}}-1}\sum_{j=0}^{i-1}\rho_{\text{C}}^j}{\rho_{\text{C}}^{n_{\text{I}}}}. \tag{3.25}$$

For further discussion, we derive some properties of $H\left(\rho_{\text{C}}\right)$.

Lemma 3.22 *If there exists a $q^* \in (0,1)$ such that $R_{\text{U}} = c_{\text{U}}W\left(q^*\right)$, it must be that $H\left(\rho_{\text{C}}\right) > 0$. Moreover, $H\left(\rho_{\text{C}}\right) > 0$ if and only if $\sum_{i=0}^{n_{\text{I}}-1}(i+1)\rho_{\text{C}}^i / \sum_{i=0}^{n_{\text{I}}-1}\rho_{\text{C}}^i < \nu_{\text{U}}$ and $H\left(\rho_{\text{C}}\right)$ strictly decreases in ρ_{C} when $H\left(\rho_{\text{C}}\right) > 0$.*

Proof of Lemma 3.22. Consider (3.24) as a quadratic equation in $\left(1-\rho_{\text{C}}+\gamma\rho\right)$.

- If $(\nu_{\mathrm{U}} - n_{\mathrm{I}})^2 + 4H(\rho_{\mathrm{c}}) < 0$, (3.24) has no real roots.
- If $(\nu_{\mathrm{U}} - n_{\mathrm{I}})^2 + 4H(\rho_{\mathrm{c}}) \geq 0$ and $H(\rho_{\mathrm{c}}) < 0$, we must have $\nu_{\mathrm{U}} - n_{\mathrm{I}} < 0$ by (3.25) and both roots of (3.24) $\frac{-(\nu_{\mathrm{U}}-n_{\mathrm{I}})\pm\sqrt{(\nu_{\mathrm{U}}-n_{\mathrm{I}})^2+4H(\rho_{\mathrm{c}})}}{2H(\rho_{\mathrm{c}})}$ are negative.
- If $(\nu_{\mathrm{U}} - n_{\mathrm{I}})^2 + 4H(\rho_{\mathrm{c}}) \geq 0$ and $H(\rho_{\mathrm{c}}) = 0$, we must have $\nu_{\mathrm{U}} - n_{\mathrm{I}} < 0$ by (3.25) and (3.24) only has one negative root $1 - \rho_{\mathrm{c}} + \gamma\rho = \frac{1}{\nu_{\mathrm{U}}-n_{\mathrm{I}}} < 0$, which is invalid because $1 - \rho_{\mathrm{c}} + \gamma\rho > 0$.
- If $H(\rho_{\mathrm{c}}) > 0$, which also implies $(\nu_{\mathrm{U}} - n_{\mathrm{I}})^2 + 4H(\rho_{\mathrm{c}}) \geq 0$, (3.24) has one positive root $\frac{-(\nu_{\mathrm{U}}-n_{\mathrm{I}})+\sqrt{(\nu_{\mathrm{U}}-n_{\mathrm{I}})^2+4H(\rho_{\mathrm{c}})}}{2H(\rho_{\mathrm{c}})}$.

Therefore, if there exists a $q^* \in (0,1)$ such that $R_{\mathrm{U}} = c_{\mathrm{U}} W(q^*)$, it must be the last case.

We next consider the monotonicity of $H(\rho_{\mathrm{c}})$. Since $\sum_{i=1}^{n_{\mathrm{I}}-1}\sum_{j=0}^{i-1}\rho_{\mathrm{c}}^j = \sum_{i=0}^{n_{\mathrm{I}}-1}(n_{\mathrm{I}}-1-i)\rho_{\mathrm{c}}^i$, rewrite $H(\rho_{\mathrm{c}})$ in an alternative form

$$H(\rho_{\mathrm{c}}) = \left(\nu_{\mathrm{U}} - \frac{\sum_{i=0}^{n_{\mathrm{I}}-1}(i+1)\rho_{\mathrm{c}}^i}{\sum_{i=0}^{n_{\mathrm{I}}-1}\rho_{\mathrm{c}}^i}\right)\frac{\sum_{i=0}^{n_{\mathrm{I}}-1}\rho_{\mathrm{c}}^i}{\rho_{\mathrm{c}}^{n_{\mathrm{I}}}}. \tag{3.26}$$

Clearly, $\sum_{i=0}^{n_{\mathrm{I}}-1}\rho_{\mathrm{c}}^i/\rho_{\mathrm{c}}^{n_{\mathrm{I}}}$ is positive and strictly decreasing in ρ_{c}. By (3.26), $H(\rho_{\mathrm{c}}) > 0 \Leftrightarrow \sum_{i=0}^{n_{\mathrm{I}}-1}(i+1)\rho_{\mathrm{c}}^i / \sum_{i=0}^{n_{\mathrm{I}}-1}\rho_{\mathrm{c}}^i < \nu_{\mathrm{U}}$. Moreover,

$$\begin{aligned}
\frac{\sum_{i=1}^{n_{\mathrm{I}}-1}(i+1)\rho_{\mathrm{c}}^i}{\sum_{i=0}^{n-1}\rho_{\mathrm{c}}^i} &= n_{\mathrm{I}} - \frac{\sum_{i=1}^{n_{\mathrm{I}}-1}(\rho_{\mathrm{c}}^i - 1) + \rho_{\mathrm{c}} - 1}{\rho_{\mathrm{c}}^n - 1} \\
&= n_{\mathrm{I}} - \frac{\sum_{i=1}^{n_{\mathrm{I}}-1}\sum_{j=0}^{i-1}\rho_{\mathrm{c}}^j + 1}{\sum_{i=0}^{n_{\mathrm{I}}-1}\rho_{\mathrm{c}}^i} \\
&= n_{\mathrm{I}} - \sum_{i=1}^{n_{\mathrm{I}}-1}\left(1 - \frac{\sum_{j=i}^{n_{\mathrm{I}}-1}\rho_{\mathrm{c}}^j}{\sum_{j=0}^{n_{\mathrm{I}}-1}\rho_{\mathrm{c}}^j}\right) - \frac{1}{\sum_{i=0}^{n_{\mathrm{I}}-1}\rho_{\mathrm{c}}^i} \\
&= n_{\mathrm{I}} - \sum_{i=1}^{n_{\mathrm{I}}-1}\left(1 - \frac{\sum_{j=0}^{n_{\mathrm{I}}-i-1}\rho_{\mathrm{c}}^j}{\sum_{j=0}^{i-1}\rho_{\mathrm{c}}^{j-i} + \sum_{j=0}^{n_{\mathrm{I}}-i-1}\rho_{\mathrm{c}}^j}\right) - \frac{1}{\sum_{i=0}^{n_{\mathrm{I}}-1}\rho_{\mathrm{c}}^i} \\
&= n_{\mathrm{I}} - \sum_{i=1}^{n_{\mathrm{I}}-1}\left(1 - \frac{1}{\frac{\sum_{j=0}^{i-1}\rho_{\mathrm{c}}^{j-i}}{\sum_{j=0}^{n_{\mathrm{I}}-i-1}\rho_{\mathrm{c}}^j} + 1}\right) - \frac{1}{\sum_{i=0}^{n_{\mathrm{I}}-1}\rho_{\mathrm{c}}^i},
\end{aligned} \tag{3.27}$$

which is strictly increasing in ρ_{c}. Consequently, we have that $H(\rho_{\mathrm{c}})$ strictly decreases in ρ_{c} when $\sum_{i=0}^{n_{\mathrm{I}}-1}(i+1)\rho_{\mathrm{c}}^i / \sum_{i=0}^{n_{\mathrm{I}}-1}\rho_{\mathrm{c}}^i < \nu_{\mathrm{U}}$, which is equivalent to $H(\rho_{\mathrm{c}}) > 0$.

■

Solving (3.24), we obtain γ as a function of ρ_{c}, i.e.,

$$\gamma\left(\rho_{\mathrm{c}}\right)=\frac{1}{\rho}\left(2\left(\left(\nu_{\mathrm{U}}-n_{\mathrm{I}}\right)+\sqrt{\left(\nu_{\mathrm{U}}-n_{\mathrm{I}}\right)^{2}+4H\left(\rho_{\mathrm{c}}\right)}\right)^{-1}+\rho_{\mathrm{c}}-1\right).$$

Since we have shown that $H\left(\rho_{\mathrm{c}}\right)$ strictly decreases in ρ_{c} when $H\left(\rho_{\mathrm{c}}\right)>0$ in Lemma 3.22, $\gamma\left(\rho_{\mathrm{c}}\right)$ then strictly increases in ρ_{c}, i.e., $\frac{d\gamma}{d\rho_{\mathrm{c}}}>0$, which implies $\frac{d\rho_{\mathrm{c}}}{d\gamma}>0$.

Step (ii). We now show that $\frac{dp_0}{d\rho_{\mathrm{c}}}<0$. Write $p_0\left(q^*\right)$ as a function of ρ_{c}:

$$\begin{aligned}
p_0\left(q^*\right) &= \left(\frac{\rho_{\mathrm{c}}^{n_{\mathrm{I}}}-1}{\rho_{\mathrm{c}}-1}+\frac{\rho_{\mathrm{c}}^{n_{\mathrm{I}}}}{1-\rho_{\mathrm{c}}+\gamma\rho}\right)^{-1}\\
&= \left(\frac{\rho_{\mathrm{c}}^{n_{\mathrm{I}}}-1}{\rho_{\mathrm{c}}-1}+\frac{\rho_{\mathrm{c}}^{n_{\mathrm{I}}}}{2}\left(\left(\nu_{\mathrm{U}}-n_{\mathrm{I}}\right)+\sqrt{\left(\nu_{\mathrm{U}}-n_{\mathrm{I}}\right)^{2}+4H\left(\rho_{\mathrm{c}}\right)}\right)\right)^{-1}\\
&= \left(\frac{\rho_{\mathrm{c}}^{n_{\mathrm{I}}}-1}{\rho_{\mathrm{c}}-1}+\frac{1}{2}\rho_{\mathrm{c}}^{n_{\mathrm{I}}}\left(\nu_{\mathrm{U}}-n_{\mathrm{I}}\right)+\sqrt{\frac{1}{4}\rho_{\mathrm{c}}^{2n_{\mathrm{I}}}\left(\nu_{\mathrm{U}}-n_{\mathrm{I}}\right)^{2}+\rho_{\mathrm{c}}^{2n_{\mathrm{I}}}H\left(\rho_{\mathrm{c}}\right)}\right)^{-1}\\
&= \left(\sum_{i=0}^{n_{\mathrm{I}}-1}\rho_{\mathrm{c}}^{i}+\frac{1}{2}\rho_{\mathrm{c}}^{n_{\mathrm{I}}}\left(\nu_{\mathrm{U}}-n_{\mathrm{I}}\right)+\sqrt{\frac{1}{4}\rho_{\mathrm{c}}^{2n_{\mathrm{I}}}\left(\nu_{\mathrm{U}}-n_{\mathrm{I}}\right)^{2}+\nu_{\mathrm{U}}\rho_{\mathrm{c}}^{n_{\mathrm{I}}}\sum_{i=0}^{n_{\mathrm{I}}-1}\rho^{i}-\rho_{\mathrm{c}}^{n_{\mathrm{I}}}\sum_{i=0}^{n_{\mathrm{I}}-1}\left(i+1\right)\rho_{\mathrm{c}}^{i}}\right)^{-1}\\
&= \left(\frac{1}{2}\rho_{\mathrm{c}}^{n_{\mathrm{I}}}\left(\nu_{\mathrm{U}}-n_{\mathrm{I}}\right)+\sum_{i=0}^{n_{\mathrm{I}}-1}\rho_{\mathrm{c}}^{i}+\sqrt{\left(\frac{1}{2}\rho_{\mathrm{c}}^{n_{\mathrm{I}}}\left(\nu_{\mathrm{U}}-n_{\mathrm{I}}\right)+\sum_{i=0}^{n_{\mathrm{I}}-1}\rho_{\mathrm{c}}^{i}\right)^{2}-\sum_{i=0}^{n_{\mathrm{I}}-1}\left(i+1\right)\rho_{\mathrm{c}}^{i}}\right)^{-1}. \qquad (3.28)
\end{aligned}$$

We notice that $\frac{1}{2}\rho_{\mathrm{c}}^{n_{\mathrm{I}}}\left(\nu_{\mathrm{U}}-n_{\mathrm{I}}\right)+\sum_{i=0}^{n_{\mathrm{I}}-1}\rho_{\mathrm{c}}^{i}>0$ when $H\left(\rho_{\mathrm{c}}\right)>0$. To see this, take the derivative in ρ_{c},

$$\begin{aligned}
&\left(\frac{1}{2}\rho_{\mathrm{c}}^{n_{\mathrm{I}}}\left(\nu_{\mathrm{U}}-n_{\mathrm{I}}\right)+\sum_{i=0}^{n_{\mathrm{I}}-1}\rho_{\mathrm{c}}^{i}\right)'\\
= &\frac{1}{2}n_{\mathrm{I}}\left(\nu_{\mathrm{U}}-n_{\mathrm{I}}\right)\rho_{\mathrm{c}}^{n_{\mathrm{I}}-1}+\frac{n_{\mathrm{I}}\rho_{\mathrm{c}}^{n_{\mathrm{I}}}-\rho_{\mathrm{c}}^{n_{\mathrm{I}}}-n_{\mathrm{I}}\rho_{\mathrm{c}}^{n_{\mathrm{I}}-1}+1}{\left(\rho_{\mathrm{c}}-1\right)^{2}}\\
> &\frac{1}{2}n_{\mathrm{I}}\left(\frac{\sum_{i=0}^{n_{\mathrm{I}}-1}\left(i+1\right)\rho_{\mathrm{c}}^{i}}{\sum_{i=0}^{n_{\mathrm{I}}-1}\rho_{\mathrm{c}}^{i}}-n_{\mathrm{I}}\right)\rho_{\mathrm{c}}^{n_{\mathrm{I}}-1}+\frac{n_{\mathrm{I}}\rho_{\mathrm{c}}^{n_{\mathrm{I}}}-\rho_{\mathrm{c}}^{n_{\mathrm{I}}}-n_{\mathrm{I}}\rho_{\mathrm{c}}^{n_{\mathrm{I}}-1}+1}{\left(\rho_{\mathrm{c}}-1\right)^{2}}\\
= &\frac{1}{2}n_{\mathrm{I}}\frac{n_{\mathrm{I}}\rho_{\mathrm{c}}-n_{\mathrm{I}}-\rho_{\mathrm{c}}^{n_{\mathrm{I}}}+1}{\left(\rho_{\mathrm{c}}-1\right)\left(\rho_{\mathrm{c}}^{n_{\mathrm{I}}}-1\right)}\rho_{\mathrm{c}}^{n_{\mathrm{I}}-1}+\frac{n_{\mathrm{I}}\rho_{\mathrm{c}}^{n_{\mathrm{I}}}-\rho_{\mathrm{c}}^{n_{\mathrm{I}}}-n_{\mathrm{I}}\rho_{\mathrm{c}}^{n_{\mathrm{I}}-1}+1}{\left(\rho_{\mathrm{c}}-1\right)^{2}}\\
= &\frac{\frac{n_{\mathrm{I}}^{2}}{2}\rho_{\mathrm{c}}^{n_{\mathrm{I}}-1}+\frac{n_{\mathrm{I}}}{2}\rho_{\mathrm{c}}^{n_{\mathrm{I}}-1}\sum_{i=0}^{n_{\mathrm{I}}-1}\rho_{\mathrm{c}}^{i}-\left(\sum_{i=0}^{n_{\mathrm{I}}-1}\rho_{\mathrm{c}}^{i}\right)^{2}}{\rho_{\mathrm{c}}^{n_{\mathrm{I}}}-1}\\
= &\frac{\frac{\left(n_{\mathrm{I}}-1\right)n_{\mathrm{I}}}{2}\rho_{\mathrm{c}}^{n_{\mathrm{I}}-1}-\sum_{i=0}^{n_{\mathrm{I}}-2}\left(i+1\right)\rho_{\mathrm{c}}^{i}+\frac{n_{\mathrm{I}}}{2}\rho_{\mathrm{c}}^{n_{\mathrm{I}}-1}\sum_{i=1}^{n_{\mathrm{I}}-1}\rho_{\mathrm{c}}^{i}-\rho_{\mathrm{c}}^{n_{\mathrm{I}}-1}\sum_{i=1}^{n_{\mathrm{I}}-1}\left(n_{\mathrm{I}}-i\right)\rho_{\mathrm{c}}^{i}}{\rho_{\mathrm{c}}^{n_{\mathrm{I}}}-1}\\
= &\frac{\frac{\left(n_{\mathrm{I}}-1\right)n_{\mathrm{I}}}{2}\rho_{\mathrm{c}}^{n_{\mathrm{I}}-1}-\sum_{i=0}^{n_{\mathrm{I}}-2}\left(i+1\right)\rho_{\mathrm{c}}^{i}}{\rho_{\mathrm{c}}^{n_{\mathrm{I}}}-1}+\frac{\rho_{\mathrm{c}}^{n_{\mathrm{I}}-1}}{2\left(\rho_{\mathrm{c}}^{n_{\mathrm{I}}}-1\right)}\sum_{i=1}^{n_{\mathrm{I}}-1}\left(2i-n_{\mathrm{I}}\right)\rho_{\mathrm{c}}^{i}
\end{aligned}$$

$$= \frac{\frac{(n_{\text{I}}-1)n_{\text{I}}}{2}\rho_{\text{c}}^{n_{\text{I}}-1} - \sum_{i=0}^{n_{\text{I}}-2}(i+1)\,\rho_{\text{c}}^{i}}{\rho_{\text{c}}^{n_{\text{I}}}-1} + \frac{\rho_{\text{c}}^{n_{\text{I}}-1}}{2\left(\rho_{\text{c}}^{n_{\text{I}}}-1\right)} \sum_{i=\left\lfloor \frac{n_{\text{I}}}{2} \right\rfloor+1}^{n_{\text{I}}-1} (2i-n_{\text{I}})\left(\rho_{\text{c}}^{2i-n_{\text{I}}}-1\right)\rho_{\text{c}}^{n_{\text{I}}-i}$$

$$> 0,$$

where the first inequality results from the fact that $\sum_{i=0}^{n_{\text{I}}-1}(i+1)\,\rho_{\text{c}}^{i}/\sum_{i=0}^{n_{\text{I}}-1}\rho_{\text{c}}^{i} < \nu_{\text{U}}$ by Lemma 3.22 and the last inequality stems from $\rho_{\text{c}} \geq 0$, which is implied by the monotonicity of $H(\rho_{\text{c}})$ and $H(\rho_{\text{c}}) > 0$. Since $\frac{1}{2}\rho_{\text{c}}^{n_{\text{I}}}(\nu_{\text{U}}-n_{\text{I}}) + \sum_{i=0}^{n_{\text{I}}-1}\rho_{\text{c}}^{i} = 1$ at $\rho_{\text{c}} = 0$. By the monotonicity, $\frac{1}{2}\rho_{\text{c}}^{n_{\text{I}}}(\nu_{\text{U}}-n_{\text{I}}) + \sum_{i=0}^{n_{\text{I}}-1}\rho_{\text{c}}^{i} > 0$ for $\rho_{\text{c}} \geq 0$, i.e., $H(\rho_{\text{c}}) > 0$.

Given the positiveness of $\frac{1}{2}\rho_{\text{c}}^{n_{\text{I}}}(\nu_{\text{U}}-n_{\text{I}}) + \sum_{i=0}^{n_{\text{I}}-1}\rho_{\text{c}}^{i}$ when $H(\rho_{\text{c}}) > 0$, for the ease of exposition, let

$$f = \left(\frac{1}{2}(\nu_{\text{U}}-n_{\text{I}})\,\rho_{\text{c}}^{n_{\text{I}}} + \sum_{i=0}^{n_{\text{I}}-1}\rho_{\text{c}}^{i}\right)^2 = \left(\frac{n_{\text{I}}\rho_{\text{c}}^{n_{\text{I}}+1} - \nu_{\text{U}}\rho_{\text{c}}^{n_{\text{I}}+1} - n_{\text{I}}\rho_{\text{c}}^{n_{\text{I}}} + \nu_{\text{U}}\rho_{\text{c}}^{n_{\text{I}}} - 2\rho_{\text{c}}^{n_{\text{I}}} + 2}{2(\rho_{\text{c}}-1)}\right)^2 \quad (3.29)$$

and

$$g = \sum_{i=0}^{n_{\text{I}}-1}(i+1)\,\rho_{\text{c}}^{i} = \frac{n_{\text{I}}\rho_{\text{c}}^{n_{\text{I}}+1} - (n_{\text{I}}+1)\,\rho_{\text{c}}^{n_{\text{I}}} + 1}{(\rho_{\text{c}}-1)^2}. \quad (3.30)$$

By (3.28), we can write $p_0(q^*) = \left(\sqrt{f} + \sqrt{f-g}\right)^{-1}$. To prove $p_0(q^*)$ is strictly decreasing in ρ_{c}, it is sufficient to show that $\sqrt{f} + \sqrt{f-g}$ is a strictly increasing function, i.e.,

$$\frac{f'}{\sqrt{f}} + \frac{f'-g'}{\sqrt{f-g}} = \frac{f'}{\sqrt{f}} - \frac{g'-f'}{\sqrt{f-g}} > 0.$$

Apparently, the inequality holds for $f' \geq g'$. We next consider the case $g' > f'$.

Note that f is strictly increasing in ν_{U} and g is independent of ν_{U}. One can also readily show that $\frac{f'}{\sqrt{f}}$ and $-\frac{g'-f'}{\sqrt{f-g}}$ are both strictly increasing in ν_{U}. Thus, if $\frac{f'}{\sqrt{f}} - \frac{g'-f'}{\sqrt{f-g}} > 0$ for $\nu_{\text{U}} = \sum_{i=0}^{n_{\text{I}}-1}(i+1)\,\rho_{\text{c}}^{i}/\sum_{i=0}^{n_{\text{I}}-1}\rho_{\text{c}}^{i}$, it must be true for all $\sum_{i=0}^{n_{\text{I}}-1}(i+1)\,\rho_{\text{c}}^{i}/\sum_{i=0}^{n_{\text{I}}-1}\rho_{\text{c}}^{i} < \nu_{\text{U}}$ by the monotonicity. As a result, $p_0(q^*)$ will be strictly decreasing in ρ_{c}, i.e., $\frac{dp_0}{d\rho_{\text{c}}} < 0$.

By the above argument, we only need to justify that $\frac{f'}{\sqrt{f}} - \frac{g'-f'}{\sqrt{f-g}} > 0$ for $\nu_{\text{U}} = \sum_{i=0}^{n_{\text{I}}-1}(i+1)\,\rho_{\text{c}}^{i}/\sum_{i=0}^{n_{\text{I}}-1}\rho_{\text{c}}^{i}$ to complete the proof. Note that

$$g' = \frac{n_{\text{I}}^2\rho_{\text{c}}^{n_{\text{I}}-1} + n_{\text{I}}\rho_{\text{c}}^{n_{\text{I}}-1} - 2\sum_{i=0}^{n_{\text{I}}-1}(i+1)\,\rho_{\text{c}}^{i}}{(\rho_{\text{c}}-1)}.$$

Moreover, at $\nu_{\mathrm{U}} = \sum_{i=0}^{n_{\mathrm{I}}-1}(i+1)\rho_{\mathrm{C}}^i / \sum_{i=0}^{n_{\mathrm{I}}-1}\rho_{\mathrm{C}}^i$,

$$f = \left(\frac{\rho_{\mathrm{C}}^{2n_{\mathrm{I}}} + n_{\mathrm{I}}\rho_{\mathrm{C}}^{n_{\mathrm{I}}+1} - 3\rho_{\mathrm{C}}^{n_{\mathrm{I}}} - n_{\mathrm{I}}\rho_{\mathrm{C}}^{n_{\mathrm{I}}} + 2}{2(\rho_{\mathrm{C}}-1)(\rho_{\mathrm{C}}^{n_{\mathrm{I}}}-1)}\right)^2,$$

and

$$f' = \frac{(n_{\mathrm{I}}\rho_{\mathrm{C}}^{n_{\mathrm{I}}+1} - 3\rho_{\mathrm{C}}^{n_{\mathrm{I}}} + \rho_{\mathrm{C}}^{2n_{\mathrm{I}}} - n_{\mathrm{I}}\rho_{\mathrm{C}}^{n_{\mathrm{I}}} + 2)\left(n_{\mathrm{I}}\rho_{\mathrm{C}}^{n_{\mathrm{I}}-1}\sum_{i=0}^{n_{\mathrm{I}}-1}\rho_{\mathrm{C}}^i - 2\left(\sum_{i=0}^{n_{\mathrm{I}}-1}\rho_{\mathrm{C}}^i\right)^2 + n_{\mathrm{I}}^2\rho_{\mathrm{C}}^{n_{\mathrm{I}}-1}\right)}{2(\rho_{\mathrm{C}}-1)(\rho_{\mathrm{C}}^{n_{\mathrm{I}}}-1)^2}.$$

Thus, evaluated at $\nu_{\mathrm{U}} = \sum_{i=0}^{n_{\mathrm{I}}-1}(i+1)\rho_{\mathrm{C}}^i / \sum_{i=0}^{n_{\mathrm{I}}-1}\rho_{\mathrm{C}}^i$,

$$\begin{aligned}
\frac{f'}{\sqrt{f}} - \frac{g'-f'}{\sqrt{f-g}} &= \frac{2}{(\rho_{\mathrm{C}}-1)^2}\frac{(\rho_{\mathrm{C}}^{n_{\mathrm{I}}}-1)}{\left(\sum_{i=0}^{n_{\mathrm{I}}-1}\rho_{\mathrm{C}}^i - n_{\mathrm{I}}\right)}\left(n_{\mathrm{I}}\rho_{\mathrm{C}}^{n_{\mathrm{I}}-1} - \sum_{i=0}^{n_{\mathrm{I}}-1}\rho_{\mathrm{C}}^i + n_{\mathrm{I}} - \sum_{i=0}^{n_{\mathrm{I}}-1}\rho_{\mathrm{C}}^i\right) \\
&= \frac{2}{(\rho_{\mathrm{C}}-1)^2}\frac{(\rho_{\mathrm{C}}^{n_{\mathrm{I}}}-1)}{\left(\sum_{i=0}^{n_{\mathrm{I}}-1}\rho_{\mathrm{C}}^i - n_{\mathrm{I}}\right)}(\rho_{\mathrm{C}}-1)\sum_{i=0}^{n_{\mathrm{I}}-2}(2i+2-n_{\mathrm{I}})\rho_{\mathrm{C}}^i \\
&= \frac{2}{(\rho_{\mathrm{C}}-1)^2}\frac{(\rho_{\mathrm{C}}^{n_{\mathrm{I}}}-1)}{\left(\sum_{i=0}^{n_{\mathrm{I}}-1}\rho_{\mathrm{C}}^i - n_{\mathrm{I}}\right)}(\rho_{\mathrm{C}}-1)\sum_{i=\lfloor\frac{n_{\mathrm{I}}}{2}\rfloor+1}^{n_{\mathrm{I}}-2}(2i+2-n_{\mathrm{I}})\left(\rho_{\mathrm{C}}^{2i+2-n_{\mathrm{I}}}-1\right)\rho_{\mathrm{C}}^{n_{\mathrm{I}}-i-2} \\
&> 0.
\end{aligned}$$

■

Proof of Theorem 3.15. The social welfare for each customer segment is

$$S_{\mathrm{I}}(q^*) = \left[\sum_{i=0}^{n_{\mathrm{I}}-1} p_i(q^*)\left(R_{\mathrm{I}} - c_{\mathrm{I}}\frac{i+1}{\mu}\right)\right]\cdot\gamma\Lambda \text{ and } S_{\mathrm{U}}(q^*) = \left[q^*\sum_{i=0}^{\infty} p_i(q^*)\left(R_{\mathrm{U}} - c_{\mathrm{U}}\frac{i+1}{\mu}\right)\right]\cdot(1-\gamma)\Lambda$$

Analogous to Theorem 4, we discuss the following cases in order: $q^* = 0$, $q^* \in (0,1)$, and $q^* = 1$.

When $q^* = 0$, $\rho_{\mathrm{C}} = \gamma\rho$. Since uninformed customers do not join, $S_{\mathrm{U}}(q^*) = 0$ and total social welfare is identical to informed individuals' contribution

$$\begin{aligned}
S_{\mathrm{I}}(q^*=0) &= \left[\sum_{i=0}^{n_{\mathrm{I}}-1} p_i(0)\left(R_{\mathrm{I}} - c_{\mathrm{I}}\frac{i+1}{\mu}\right)\right]\cdot\gamma\Lambda \\
&= \left(\frac{1-(\gamma\rho)^{n_{\mathrm{I}}}}{1-\gamma\rho} + (\gamma\rho)^{n_{\mathrm{I}}}\right)^{-1}\left(R_{\mathrm{I}}\frac{1-(\gamma\rho)^{n_{\mathrm{I}}}}{1-\gamma\rho} - \frac{c_{\mathrm{I}}}{\mu}\frac{1-(n_{\mathrm{I}}+1)(\gamma\rho)^{n_{\mathrm{I}}}+n(\gamma\rho)^{n_{\mathrm{I}}+1}}{(1-\gamma\rho)^2}\right)\cdot\gamma\Lambda.
\end{aligned}$$

Notice that $S_{\mathrm{I}}(q^*=0)$ is independent of R_{U}. Thus, we can apply the same discussion in the proof of Theorem 4(i) to show that $S_{\mathrm{I}}(q^*=0) + S_{\mathrm{U}}(q^*=0)$ strictly decreases in γ.

When $q^* \in (0,1)$, the social welfare yielded by uninformed customers equals zero as

well, i.e., $S_{\text{U}}(q^*) = 0$. Thus, we only need to consider $S_{\text{I}}(q^*)$.

$$\begin{aligned}
S_{\text{I}}(q^*) &= \left[\sum_{i=0}^{n_{\text{I}}-1} p_i(q^*)\left(R_{\text{I}} - c_{\text{I}}\frac{i+1}{\mu}\right)\right]\cdot\gamma\Lambda \\
&= p_0(q^*)\left(R_{\text{I}}\frac{1-\rho_{\text{c}}^{n_{\text{I}}}}{1-\rho_{\text{c}}} - \frac{c_{\text{I}}}{\mu}\frac{1-(n_{\text{I}}+1)\rho_{\text{c}}^{n_{\text{I}}}+n\rho_{\text{c}}^{n_{\text{I}}+1}}{(1-\rho_{\text{c}})^2}\right)\cdot\gamma\Lambda \\
&= c_{\text{I}}p_0(q^*)\left(\nu_{\text{I}}\frac{1-\rho_{\text{c}}^{n_{\text{I}}}}{1-\rho_{\text{c}}} - \frac{1-(n_{\text{I}}+1)\rho_{\text{c}}^{n_{\text{I}}}+n_{\text{I}}\rho_{\text{c}}^{n_{\text{I}}+1}}{(1-\rho_{\text{c}})^2}\right)\cdot\rho\gamma(\rho_{\text{c}}) \\
&= c_{\text{I}}\rho_{\text{c}}^{n_{\text{I}}}\left(H(\rho_{\text{c}}) - \frac{(\nu_{\text{U}}-\nu_{\text{I}})\sum_{i=0}^{n_{\text{I}}-1}\rho_{\text{c}}^i}{\rho_{\text{c}}^{n_{\text{I}}}}\right)\cdot p_0(q^*)\cdot\rho\gamma(\rho_{\text{c}}) \\
&= c_{\text{I}}\rho_{\text{c}}^{n_{\text{I}}}H(\rho_{\text{c}})\cdot p_0(q^*)\cdot\rho\gamma(\rho_{\text{c}}) - c_{\text{I}}(\nu_{\text{U}}-\nu_{\text{I}})\sum_{i=0}^{n_{\text{I}}-1}\rho_{\text{c}}^i\cdot p_0(q^*)\cdot\rho\gamma(\rho_{\text{c}}) \\
&= c_{\text{I}}\left(1-(\nu_{\text{U}}-\nu_{\text{I}})\frac{\sum_{i=0}^{n_{\text{I}}-1}\rho_{\text{c}}^i}{\rho_{\text{c}}^{n_{\text{I}}}H(\rho_{\text{c}})}\right)\rho_{\text{c}}^{n_{\text{I}}}H(\rho_{\text{c}})\cdot p_0(q^*)\cdot\rho\gamma(\rho_{\text{c}}) \\
&= c_{\text{I}}\left[1-(\nu_{\text{U}}-\nu_{\text{I}})\left(\nu_{\text{U}} - \frac{\sum_{i=0}^{n_{\text{I}}-1}(i+1)\rho_{\text{c}}^i}{\sum_{i=0}^{n_{\text{I}}-1}\rho_{\text{c}}^i}\right)^{-1}\right]\rho_{\text{c}}^{n_{\text{I}}}H(\rho_{\text{c}})\cdot p_0(q^*)\cdot\rho\gamma(\rho_{\text{c}}) \qquad (3.31)
\end{aligned}$$

We first observe that $\rho_{\text{c}}^{n_{\text{I}}}H(\rho_{\text{c}})\cdot p_0(q^*)\cdot\rho\gamma(\rho_{\text{c}})$ strictly increases in ρ_{c}. Substitute $p_0(q^*)$ with (3.28),

$$\begin{aligned}
&\rho_{\text{c}}^{n_{\text{I}}}H(\rho_{\text{c}})\,p_0(q^*)\,\rho\gamma(\rho_{\text{c}}) \\
=\ &\rho_{\text{c}}^{n_{\text{I}}}H(\rho_{\text{c}})\frac{2\left((\nu_{\text{U}}-n_{\text{I}})+\sqrt{(\nu_{\text{U}}-n_{\text{I}})^2+4H(\rho_{\text{c}})}\right)^{-1}+\rho_{\text{c}}-1}{\frac{1-\rho_{\text{c}}^{n_{\text{I}}}}{1-\rho_{\text{c}}}+\frac{\rho_{\text{c}}^{n_{\text{I}}}}{2}\left((\nu_{\text{U}}-n_{\text{I}})+\sqrt{(\nu_{\text{U}}-n_{\text{I}})^2+4H(\rho_{\text{c}})}\right)} \\
=\ &\left(2\rho_{\text{c}}^{n_{\text{I}}}H(\rho_{\text{c}})+(\nu_{\text{U}}-n_{\text{I}})\rho_{\text{c}}^{n_{\text{I}}}H(\rho_{\text{c}})(\rho_{\text{c}}-1)+\rho_{\text{c}}^{n_{\text{I}}}H(\rho_{\text{c}})(\rho_{\text{c}}-1)\sqrt{(\nu_{\text{U}}-n_{\text{I}})^2+4H(\rho_{\text{c}})}\right)\times \\
&\left((\nu_{\text{U}}-n_{\text{I}})\frac{1-\rho_{\text{c}}^{n_{\text{I}}}}{1-\rho_{\text{c}}}+(\nu_{\text{U}}-n_{\text{I}})^2\rho_{\text{c}}^{n_{\text{I}}}+2\rho_{\text{c}}^{n_{\text{I}}}H(\rho_{\text{c}})+\right. \\
&\qquad\left.\left(\frac{1-\rho_{\text{c}}^{n_{\text{I}}}}{1-\rho_{\text{c}}}+(\nu_{\text{U}}-n_{\text{I}})\rho_{\text{c}}^{n_{\text{I}}}\right)\sqrt{(\nu_{\text{U}}-n_{\text{I}})^2+4H(\rho_{\text{c}})}\right)^{-1} \\
=\ &\left(1-\nu_{\text{U}}\frac{(\nu_{\text{U}}-n_{\text{I}})+\sqrt{(\nu_{\text{U}}-n_{\text{I}})^2+4H(\rho_{\text{c}})}}{2\rho_{\text{c}}^{n_{\text{I}}}H(\rho_{\text{c}})+\left(\frac{1-\rho_{\text{c}}^{n_{\text{I}}}}{1-\rho_{\text{c}}}+(\nu_{\text{U}}-n_{\text{I}})\rho_{\text{c}}^{n_{\text{I}}}\right)\left((\nu_{\text{U}}-n_{\text{I}})+\sqrt{(\nu_{\text{U}}-n_{\text{I}})^2+4H(\rho_{\text{c}})}\right)}\right)
\end{aligned}$$

$$= \left(1 - \frac{\nu_{\text{U}}}{\frac{1-\rho_{\text{C}}^{n_{\text{I}}}}{1-\rho_{\text{C}}} + \frac{\rho_{\text{C}}^{n_{\text{I}}}}{2}\left(\left(\nu_{\text{U}} - n_{\text{I}}\right) + \sqrt{\left(\nu_{\text{U}} - n_{\text{I}}\right)^2 + 4H\left(\rho_{\text{C}}\right)}\right)}\right)$$

$$= 1 - \nu_{\text{U}} p_0\left(q^*\right). \tag{3.32}$$

We have already demonstrated that $p_0\left(q^*\right)$ strictly decreases in ρ_{C} in the proof of Theorem 3.14. Therefore, $\rho_{\text{C}}^{n_{\text{I}}} H\left(\rho_{\text{C}}\right) \cdot p_0\left(q^*\right) \cdot \rho\gamma\left(\rho_{\text{C}}\right)$ also strictly increases in ρ_{C}.

Next, we consider the monotonicity of the term in the square bracket of (3.31). Recall that $\sum_{i=0}^{n_{\text{I}}-1}\left(i+1\right)\rho_{\text{C}}^i \big/ \sum_{i=0}^{n_{\text{I}}-1}\rho_{\text{C}}^i$ strictly increases in ρ_{C} as shown in the proof of Theorem 5. Then,

- If $\nu_{\text{U}} \leq \nu_{\text{I}}$, $1-\left(\nu_{\text{U}}-\nu_{\text{I}}\right)\left(\nu_{\text{U}} - \frac{\sum_{i=0}^{n_{\text{I}}-1}(i+1)\rho_{\text{C}}^i}{\sum_{i=0}^{n_{\text{I}}-1}\rho_{\text{C}}^i}\right)^{-1}$ is increasing in ρ_{C}. In this case, $S_{\text{I}}\left(q^*\right)$ is increasing in ρ_{C}. Due to the fact that $\frac{d\rho_{\text{C}}}{d\gamma} > 0$, we have $S_{\text{I}}\left(q^*\right)$ is increasing in γ.

- If $\nu_{\text{U}} > \nu_{\text{I}}$, $1-\left(\nu_{\text{U}}-\nu_{\text{I}}\right)\left(\nu_{\text{U}} - \frac{\sum_{i=0}^{n_{\text{I}}-1}(i+1)\rho_{\text{C}}^i}{\sum_{i=0}^{n_{\text{I}}-1}\rho_{\text{C}}^i}\right)^{-1}$ is decreasing in ρ_{C}. Then $S_{\text{I}}\left(q^*\right)$ might be unimodal in ρ_{C}, which leads to that $S_{\text{I}}\left(q^*\right)$ might be unimodal in γ.

When $q^* = 1$, $\rho_{\text{C}} = \rho$. The total social welfare is

$$\begin{aligned}
S_{\text{I}}\left(q^*\right) + S_{\text{U}}\left(q^*\right) &= \left[\sum_{i=0}^{n_{\text{I}}-1} p_i\left(1\right)\left(R_{\text{I}} - c_{\text{I}}\frac{i+1}{\mu}\right)\right]\cdot\gamma\Lambda + \left[\sum_{i=0}^{\infty} p_i\left(1\right)\left(R_{\text{U}} - c_{\text{U}}\frac{i+1}{\mu}\right)\right]\cdot\left(1-\gamma\right)\Lambda \\
&= \left[\sum_{i=0}^{n_{\text{I}}-1} p_i\left(1\right)\left(R_{\text{I}} - c_{\text{I}}\frac{i+1}{\mu}\right)\right]\cdot\Lambda + \left[\sum_{i=n_{\text{I}}}^{\infty} p_i\left(1\right)\left(R_{\text{U}} - c_{\text{U}}\frac{i+1}{\mu}\right)\right]\cdot\left(1-\gamma\right)\Lambda \\
&= \frac{c_{\text{I}}\Lambda}{\mu} p_0\left(1\right)\left(\frac{R_{\text{I}}\mu}{c_{\text{I}}}\cdot\frac{1-\rho^{n_{\text{I}}}}{1-\rho} - \frac{1-\left(n_{\text{I}}+1\right)\rho^{n_{\text{I}}}+n_{\text{I}}\rho^{n_{\text{I}}+1}}{\left(1-\rho\right)^2}\right) \\
&\qquad + \frac{c_{\text{U}}\Lambda}{\mu} p_0\left(1\right)\rho^{n_{\text{I}}}\left(\frac{R_{\text{U}}\mu}{c_{\text{U}}}\cdot\frac{1}{1-\rho+\gamma\rho} - \frac{n_{\text{I}}\left(1-\rho+\gamma\rho\right)+1}{\left(1-\rho+\gamma\rho\right)^2}\right)\left(1-\gamma\right) \\
&= \frac{c_{\text{I}}\Lambda}{\mu} p_0\left(1\right)\nu_{\text{I}}\frac{1-\rho^{n_{\text{I}}}}{1-\rho} - \frac{c_{\text{I}}\Lambda}{\mu} p_0\left(1\right)\frac{1-\left(n_{\text{I}}+1\right)\rho^{n_{\text{I}}}+n_{\text{I}}\rho^{n_{\text{I}}+1}}{\left(1-\rho\right)^2} \\
&\qquad + \frac{c_{\text{U}}\Lambda}{\mu} p_0\left(1\right)\rho^{n_{\text{I}}}\left(\frac{\nu_{\text{U}}}{1-\rho+\gamma\rho} - \frac{n_{\text{I}}\left(1-\rho+\gamma\rho\right)+1}{\left(1-\rho+\gamma\rho\right)^2}\right)\left(1-\gamma\right) \\
&= \frac{c_{\text{I}}\Lambda}{\mu} p_0\left(1\right)\left[\frac{n_{\text{I}}\left(1-\rho\right)-\left(1-\rho^{n_{\text{I}}}\right)}{\left(1-\rho\right)^2} + \left\langle\nu_{\text{I}}\right\rangle\frac{1-\rho^{n_{\text{I}}}}{1-\rho}\right] \\
&\qquad + \frac{c_{\text{U}}\Lambda}{\mu} p_0\left(1\right)\rho^{n_{\text{I}}}\left[\frac{\left(\nu_{\text{U}}-n_{\text{I}}\right)\left(1-\gamma\right)}{1-\rho+\gamma\rho} - \frac{\left(1-\gamma\right)}{\left(1-\rho+\gamma\rho\right)^2}\right] \\
&= \frac{\Lambda}{\mu} p_0\left(1\right)\left[c_{\text{I}}\frac{n_{\text{I}}\left(1-\rho\right)-\left(1-\rho^{n_{\text{I}}}\right)}{\left(1-\rho\right)^2} + c_{\text{I}}\left\langle\nu_{\text{I}}\right\rangle\frac{1-\rho^{n_{\text{I}}}}{1-\rho}\right. \\
&\qquad \left. + c_{\text{U}}\rho^{n_{\text{I}}}\frac{\left(\nu_{\text{U}}-n_{\text{I}}\right)\left(1-\gamma\right)}{1-\rho+\gamma\rho} - c_{\text{U}}\rho^{n_{\text{I}}}\frac{\left(1-\gamma\right)}{\left(1-\rho+\gamma\rho\right)^2}\right]
\end{aligned}$$

Note that $p_0(q^*=1)=\left(\frac{1-\rho^{n_{\text{I}}}}{1-\rho}+\frac{\rho^{n_{\text{I}}}}{1-\rho+\gamma\rho}\right)^{-1}$ strictly increases in γ and the term in the bracket is positive (since $S_{\text{I}}(q^*)+S_{\text{U}}(q^*)>0$ and $p_0(q^*=1)>0$). Therefore, we only need to explore the monotonicity of $\frac{(\nu_{\text{U}}-n_{\text{I}})(1-\gamma)}{1-\rho+\gamma\rho}-\frac{(1-\gamma)}{(1-\rho+\gamma\rho)^2}$ w.r.t γ. Take the derivative,

$$\frac{\partial}{\partial\gamma}\left(\frac{(\nu_{\text{U}}-n_{\text{I}})(1-\gamma)}{1-\rho+\gamma\rho}-\frac{(1-\gamma)}{(1-\rho+\gamma\rho)^2}\right)=\frac{1+\rho-\gamma\rho-(\nu_{\text{U}}-n_{\text{I}})(1-\rho+\gamma\rho)}{(1-\rho+\gamma\rho)^3}$$

Note that $(1-\gamma)\rho$ is the workload caused by uninformed customers. Due to the fact that uninformed customers join the queue with probability 1, the server must have enough capacity to handle all of them, i.e., $(1-\gamma)\rho<1\Leftrightarrow 1-\rho+\gamma\rho\geq 0$. Thus,

- If $(\nu_{\text{U}}-n_{\text{I}})\leq\frac{1+\rho-\gamma\rho}{1-\rho+\gamma\rho}$, we have $\frac{(\nu_{\text{U}}-n_{\text{I}})(1-\gamma)}{1-\rho+\gamma\rho}-\frac{(1-\gamma)}{(1-\rho+\gamma\rho)^2}$ is increasing in γ. Then, $S_{\text{I}}(q^*=1)+S_{\text{U}}(q^*=1)$ is increasing in γ.
- If $(\nu_{\text{U}}-n_{\text{I}})>\frac{1+\rho-\gamma\rho}{1-\rho+\gamma\rho}$, we have $\frac{(\nu_{\text{U}}-n_{\text{I}})(1-\gamma)}{1-\rho+\gamma\rho}-\frac{(1-\gamma)}{(1-\rho+\gamma\rho)^2}$ is decreasing in γ. In this case, $S_{\text{I}}(q^*=1)+S_{\text{U}}(q^*=1)$ might be unimodal in γ.

■

Chapter 4

Capacity Allocation under Endogenous Arrivals

4.1 Introduction

Models of congestion-prone service systems, such as call centers and hospital emergency departments, usually assume that customers arrive exogenously, specifically independent of the wait time conditions at a service facility. However, delays at facilities are often time-varying (Gans et al. 2003, Akşin et al. 2007). For example, many call centers are most congested in the late morning and/or mid afternoon, so customers are more likely to experience long service delays during these times (Brown et al. 2005, Ibrahim and L'Ecuyer 2013). The temporary burst in arrivals may reflect most customers' time-of-service (TOS) preferences. On the other hand, some customers may be flexible in TOS but more sensitive to service delays. It seems plausible to expect that these customers may change their calling times in order to avoid long waiting times. Behavioral studies based on laboratory experiments have revealed that people do adjust their arrival times according to anticipated delays at different times (Rapoport et al. 2004, Seale et al. 2005, Stein et al. 2007). The same studies also demonstrate that people appear to be heterogeneous in their sensitivities to delays.

In light of customers' tendency to adjust arrivals in exchange for shorter delays, service providers may want to exploit the opportunity to reduce demand variability by informing customers about the time windows with shorter delays. For example, St. Mary's General Hospital (SMGH) in Kitchener, Canada, posts emergency department (ED) wait time information online. From their website[1], potential patients can learn the number of

[1] http://www.smgh.ca/ed-wait-times

people waiting, the number being treated in the emergency department, the estimated current wait times, and, *in particular*, the predicted wait times over the next six hours. According to the hospital President Don Shilton, "It (predicted wait time information) helped them (patients) make informed decisions about when to come to the ED..."[2]

Beyond using delay information to influence patients' arrival process, providers also account for demand variability in their capacity decisions. This raises an important yet less studied research question: What is the best capacity allocation policy when customers' arrival time decisions account for intertemporal fluctuations in service delays? The intra-day arrivals to call centers or the SMGH ED may, to a certain extent, result from customer responses to fluctuating service delays over the day. These intra-day service delays in turn depend on the provider's intertemporal capacity decisions. Therefore, customer arrivals become endogenous, that is, they depend on the capacity allocation. Previous studies on service capacity management, however, ignore this underlying endogeneity between capacity allocation and the arrival process. In this paper, we incorporate this endogeneity and investigate the intertemporal capacity allocation problem when customer arrivals are (partially) responsive to facility delays over time.

We model customers' self-interested TOS decisions over several consecutive discrete periods. Our model allows customer heterogeneity in both TOS preferences and delay sensitivities. In contrast to other work on customer strategic arrivals that takes capacity to be fixed (Glazer and Hassin 1983, Lariviere and Van Mieghem 2004, Honnappa and Jain 2015), we focus on the interplay between customers' utility-maximizing TOS choices and the provider's intertemporal capacity decisions. We consider the provider's problem in two models: one under limited total capacity, the other with a time-varying capacity cost. The former applies to industries with inadequate supply of key resources, such as health care services, where shortage of physicians is perceived as one of the main problems (Fortune 2014) and the main concern is to efficiently utilize available capacity. The latter model adds a cost perspective to capacity management and reflects the practical reality that operating cost often oscillates over the day (e.g., Gurvich et al. 2014).

Although our model focuses on customers' intertemporal TOS choices at a single facility, it can also represent customers' server (routing) choices at a given time among multiple facilities. For example, health authorities in the city of Vancouver create an online real-time ED waiting time dashboard[3], attempting to efficiently use the region's hospital resources and improve patient flow. The system lists waiting times of five bus-

[2] http://www.smgh.ca/st-marys-real-time-wait-time-website-wins-national-innovation-prize

[3] http://www.edwaittimes.ca

iest emergency departments in the region and displays the locations of these emergency departments on the city map. Potential patients determine the ED to visit, based on both the traveling distances and the posted wait times.

We show that for any given capacity allocation, there exists a customer TOS choice equilibrium. There may be multiple TOS choice equilibria, but the total customer arrival rate to a particular period is unique. Taking into account customer equilibrium TOS choices, we then analyze the service provider's capacity allocation/investment decisions, with the objective to maximize the total system utility. We find that under the optimal capacity allocation, any TOS equilibrium is in pure strategies, unlike when capacity allocation is fixed.

In terms of the provider's ability to achieve the system optimal result through capacity control alone, we have two findings. On the one hand, we show that when the total capacity (e.g., the number of physician shifts) is fixed and customers do not balk, then the ability to adjust the capacity allocation effectively avoids system efficiency losses due to customers' self-interested decisions. On the other hand, we demonstrate that capacity management alone cannot align the incentives of the system with those of individual customers when monetary cost also plays a role, as in our time-varying cost model. In that case, a pricing scheme has to be imposed to give customers the incentive to implement the socially efficient outcome. Specifically, the provider needs to charge for TOS such that price differences across time periods equal the corresponding capacity cost differences. Moreover, setting prices equal to the capacity costs also induces the system optimal arrival rates.

Section 4.2 briefly reviews the related literature. Section 4.3 presents the model. Section 4.4 establishes existence of the customer TOS choice equilibrium. Section 4.5 analyzes the capacity allocation problems. Section 4.6 provides concluding remarks.

4.2 Literature Review

Studies of customers' rational responses to service delays have been active in service operations since the pioneer work of Naor (1969), which considers customer joining/balking decisions in the presence of congestion. Hassin and Haviv (2003) provide a comprehensive survey on numerous extensions to Naor (1969). While waiting in the line, customers may change their minds and abandon from the queue. A stream of research investigates the abandonment behavior theoretically (e.g., Mandelbaum and Shimkin 2000, Shimkin and Mandelbaum 2004) or empirically (e.g., Akşin et al. (2013)). Instead of leaving for good after balking or abandoning, customers may also try to seek service again at later

times. There is a rich literature on retrial models and their implications for operational decisions (e.g., Mandelbaum et al. 2002, Aguir et al. 2008, Cui et al. 2014, de Véricourt and Zhou 2005, Armony and Maglaras 2004a,b).

None of the aforementioned papers on customer strategic behavior explicitly model how a customer decides at which time she should visit a facility in order to shorten anticipated delays. The earliest work that explicitly considers the equilibrium arrival time decisions of customers dates back to the seminal bottleneck model of Vickrey (1969). Arnott et al. (1993) extend the analysis to elastic demand and examine its economic implications. We refer to de Palma and Fosgerau (2011), Small (2015) and references therein for a comprehensive review and discussion of dynamic traffic congestion models.

However, these traffic models typically tend to ignore the system stochasticity, which is one of the challenges in service operations. Glazer and Hassin (1983) appear to be the first to consider a stochastic model; they model strategic arrival decisions to a facility that opens at a fixed time and remains open thereafter. Customers are allowed to queue before the facility opens and aim to minimize their expected delay costs. The authors characterize the system transient behavior by deriving the customer arrival rate in equilibrium. Hassin and Kleiner (2010) reconsider the problem for the case where early arrivals are forbidden. Lariviere and Van Mieghem (2004) study a similar problem in which customers find congestion costly and seek service when the facility is underutilized. They work in discrete time and show that the equilibrium arrival process converges to a Poisson process. Juneja and Jain (2009) consider the concert game, a scenario in which all customers prefer the service as soon as the facility is open. Examples of such scenarios include line-up for concerts, games, or Black Friday sales. While early appearances incur waiting cost, late arrivals also yield opportunity cost. Juneja and Jain (2009) derive the equilibrium arrival strategy of homogeneous customers and measure the system efficiency loss due to customer self-interested decisions. Jain et al. (2011) and Honnappa and Jain (2015) extend the analysis to multiple customer classes and queueing networks, respectively.

We distinguish our work from the above papers on strategic arrivals in that we study the interaction between capacity decisions and customer choice behaviors. Prior studies all treat service capacity as a fixed constant over time, whereas we are interested in optimal intertemporal capacity allocation to achieve system efficiency while accounting for customer responses in the timing of arrivals.

Our research is also related to studies on customer server choices, because we can regard each time period as a distinct facility. In contrast to our model, many papers in this line of research assume server-independent valuations, i.e., customers are only

sensitive to delays, sometimes prices as well, but do not discriminate among servers that are actually used. Bell and Stidham (1983) study the system equilibrium under customer self-interested server choices when all service rates are given. They show that individual choices do not result in socially optimal outcomes. Recent developments on the price of anarchy in selfish routing games provide measurements on the efficiency loss in more general settings (e.g., Roughgarden and Tardos 2002, Roughgarden 2005 and references therein, Haviv and Roughgarden 2007). It has also been widely studied how *competing* firms strategically determine their control variables in the presence of customer choices. Luski (1976) and Levhari and Luski (1978) discuss firms' price decisions in duopoly while service rates are exogenous. Alternatively, Kalai et al. (1992) consider duopoly firms' capacity choices. In the selfish routing literature, Acemoglu et al. (2009) and Johari et al. (2010) allow competing providers to control both prices and capacities.

While server-independent valuations are applicable in certain settings, e.g., selfish routing in communication networks, customers do have preferences on servers by which they are being served in other scenarios, i.e., service valuations can be server-dependent. Allon and Federgruen (2007) model the demand rates to oligopoly firms as aggregate functions of prices and expected delays, using attraction models and similar reduced-form customer choice models. Unlike in our model, incentive-compatibility issues are therefore absent in theirs, and they focus on competitive equilibria, rather than on system optimality. In contrast to the above papers, the work of Veeraraghavan and Debo (2009, 2011) assumes that customers are *uncertain* about the quality of (their valuation for) the servers. They characterize customers' equilibrium server choices, accounting for customer inference on product quality from queue lengths. Finally, customers' service valuations may be affected by traveling distance to facility. The literature on location models focuses on the design of service networks by selecting facility locations. We refer to Berman and Krass (2015) for an extensive review of stochastic location models with congestion. This stream of research assumes that the capacity is fixed for each location. We find that the ability to set capacity endogenously induces customers of the same type to use a pure strategy in choosing their time-of-service. This contrasts with the result in the location literature that only mixed strategy equilibria are guaranteed under self-interested decisions as illustrated in Berman and Krass (2015).

4.3 Customer Characteristics and Time-of-Service

A service provider operates a facility with a first-in first-out (FIFO) discipline over a finite time horizon, e.g., a call center that is open over a certain time window during a

day. We divide the time horizon during which the service is available into n consecutive non-overlapping short periods. For example, in call centers, a time interval of 15 or 30 minutes in length is often used for statistical purposes (Brown et al. 2005, Gans et al. 2003). For period j, $j = 1, \ldots, n$, there is a stream of base customers, who are extremely sensitive to time-of-service (TOS). They only consider service in period j valuable and are not willing to be served in other time periods. Base customers of period j visit the facility according to a Poisson process at a rate of Λ_j^b. Their service valuation is v_j^b and a linear delay cost is incurred at a rate of c_j^b per unit time. Therefore, a base customer of period j receives an expected surplus of $v_j^b - c_j^b W_j$ upon completion of the service, where W_j is the expected waiting time, including service time, in period j.

In addition to the base customers, there is also a stream of strategic customers, who may have their preferences on TOS but are also willing to consider services in alternative periods in exchange for shorter delays. We classify strategic customers by their service valuation vector and unit delay cost vector. Type i strategic customers, who request the service at a total rate of Λ_i, value period j's service v_{ij} and their unit delay cost in period j is c_{ij}. We can treat the order of v_{ij}'s as a representation of type i strategic customers intrinsic TOS preferences in the absence of delays. For instance, $v_{11} > v_{12}$ indicates that type 1 strategic customers prefer service in period 1 than that in period 2 if delays in both periods are not taken into account. Two strategic customers are of the same type if and only if their valuations and unit delay costs are identical in all periods. Therefore, $\boldsymbol{v}_i = (v_{i1}, \ldots, v_{in})$ and $\boldsymbol{c}_i = (c_{i1}, \ldots, c_{in})$ characterize type i strategic customers. We assume that there are m types of strategic customers. For ease of exposition, we sometimes omit the word "strategic" when refer to type i strategic customers if it is clear that we are not referring to base customers.

Our strategic customer model is reasonably general. It covers many common settings as special cases. For example, it can capture customer heterogeneous valuations with time-invariant delay cost or similarly heterogeneous delay costs with a constant valuation.

In contrast to base customers, who only visit specific periods, strategic customers choose their TOS, i.e., which period to join for the service. All strategic customers simultaneously choose their TOS and visit the facility at a random instant during that period. Hence, their TOS decisions form an arrival rate matrix $\boldsymbol{\Lambda}$, where Λ_{ij} represents the arrival rate of type i customers who choose to join period j. As a result, total arrivals to period j include not only base customers but also strategic customers who choose to be served in period j, i.e.,

$$\lambda_j(\boldsymbol{\Lambda}) = \Lambda_j^b + \sum_{i=1}^{m} \Lambda_{ij} \text{ for } j = 1, \ldots, n. \tag{4.1}$$

Further, let $C_j(\lambda_j)$ denote the total unit delay cost for period j at an arrival rate λ_j. Then, we have

$$C_j(\lambda_j) = c_j^b \Lambda_j^b + \sum_{i=1}^{m} c_{ij}\Lambda_{ij} \text{ for } j = 1, \ldots, n. \tag{4.2}$$

Let $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$ represent the intertemporal service rate vector. We further assume the queueing dynamic in each period as an M/M/1 system, i.e., the interarrival times and service times in period j are exponentially distributed with means $1/\lambda_j$ and $1/\mu_j$ respectively. We focus our attention on a single server system for tractability. As a first step in understanding the interaction between capacity allocation and customer intertemporal arrivals, the simplicity of an M/M/1 system enables us to analytically derive key properties of the optimal staffing and customer routing decisions.

We also make the simplifying assumption that time periods are independent. That is, we ignore the transient effect that may occur as arrival rates may change from one period to the next. Therefore, we only consider the system performances in steady states and period j's expected waiting time $W_j(\boldsymbol{\Lambda}) = (\mu_j - \lambda_j)^{-1}$, where the arrival rate matrix $\boldsymbol{\Lambda}$ and λ_j are defined in (4.1). Our model can be considered as an approximation for a non-stationary system. Previous literature (e.g., Green and Kolesar 1991, Green et al. 2001) suggests that such approximations, applying stationary models to non-stationary systems, performs well when the service times are short and the quality-of-service standard is high, such as services provided by call centers. In fact, many commercial call-center-management softwares are implemented based on these approximations (Green et al. 2001). On the other hand, such approximations may not perform well for cases with medium to long service times or systems that are overloaded for extensive period of time (Green et al. 2007). This drawback of the current model inspires future research for more sophisticated frameworks.

Upon completion of the service, a base customer of period j and a type i customer who joins period j expect to receive net utilities

$$U_j^b(\boldsymbol{\Lambda}) = v_j^b - c_j^b W_j(\boldsymbol{\Lambda}) \text{ and } U_{ij}(\boldsymbol{\Lambda}) = v_{ij} - c_{ij} W_j(\boldsymbol{\Lambda}),$$

respectively. A strategic customer's objective is to choose her arrival period or TOS that maximizes her expected utility. For convenience, we first assume strategic customers cannot balk and we shall relax this assumption in Sections 4.5.3.2 and 4.5.4.2.

Lastly, we assume all parameters of the game, namely arrival rates, service speeds, and the parameters of all customers' utility functions, are common knowledge.

As we mentioned in the introduction, our model can also be interpreted as a rep-

resentation of interspatial choices over facility locations. In the example of Vancouver hospital wait time program, EDs are spatially substitutable for patients. In this case, facility locations of hospitals, indexed by j, are parallel to the time periods in the TOS model. Base customers may represent patients sufferring from critical conditions, such as severe chest pains or breathing problems. These patients require immediate medical treatments and thus they would almost certainly visit the closest emergency room. Likewise, strategic customers represent the patients with noncritical conditions, such as ankle sprains. Although nearby emergency rooms save travel distances, long wait times at these facilities may encourage potential patients to visit other emergency rooms as long as delay information of all facilities is publicly available. Accordingly, we can define patients from the same neighborhood with the same delay sensitivity as the same type. For a type i patient, her expected utility of visiting facility at location j equals $v_{ij} - c_i W_j$, where the service valuation v_{ij} might be a function of the distance to facility location j, and the unit delay cost c_i is likely to be invariant regardless of locations. Finally, patient emergency room choices constitute arrival processes at all facilities, which can be captured by the arrival rate matrix $\mathbf{\Lambda}$. This example demonstrates how we can apply our customer choice model to represent facility substitutability with server-dependent valuations. However, for consistency, we concentrate on intertemporal choices of TOS in this paper.

4.4 Equilibrium of Time-of-Service Choices

Every strategic customer seeks service in the time period that maximizes her expected net utility $U_{ij}(\mathbf{\Lambda})$. In this TOS choice game, each strategic customer, or player, is nonatomic in the sense that there are a large number of players, each controlling a negligible fraction of the overall arrivals. For this reason, we are not interested in strategic customer individual choices, but focus on the system intertemporal arrivals at an aggregate level. Specifically, we are interested in equilibrium arrival rate matrices $\mathbf{\Lambda}^*$ that result from choices of all strategic customers and capture the system demand as a whole. We first define the TOS equilibrium and show its existence. Although there might be multiple equilibria, we prove that all equilibria yield the same arrival rate vector $\boldsymbol{\lambda}^* = (\lambda_1^*, \ldots, \lambda_n^*)$.

Definition 1 (Equilibrium of Time-of-Service choices) *Assume capacities in all periods are given, i.e., $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$ is fixed. An arrival rate matrix $\mathbf{\Lambda}^*$ is an equilibrium if and only if the following two conditions hold for any strategic customer of type i, $i = 1, \ldots, m$,*

(EC1) $v_{ij} - c_{ij}W_j(\mathbf{\Lambda}^*) = v_{ij'} - c_{ij'}W_{j'}(\mathbf{\Lambda}^*)$ *for all* j *and* j' *with* $\Lambda^*_{ij}, \Lambda^*_{ij'} > 0$, *and*
(EC2) $v_{ij} - c_{ij}W_j(\mathbf{\Lambda}^*) \geq v_{ij'} - c_{ij'}W_{j'}(\mathbf{\Lambda}^*)$ *for all* j *and* j' *with* $\Lambda^*_{ij} > 0$ *and* $\Lambda^*_{ij'} = 0$.

Our definition of TOS equilibrium adopts Wardrop's first principle in Wardrop and Whitehead (1952), which is commonly used for the prediction of traffic patterns in transportation networks. By (EC1), a strategic customer is indifferent among time periods that have been chosen by the cohort of her type. Moreover, (EC2) indicates that she also has no incentive to deviate from her current choice to a period that have not been chosen by the cohort of her type, since the deviation would not result in higher utility. In other words, in equilibrium, for strategic customers of the same type, expected net utilities in all chosen periods are equal, and no less than those in any unchosen period. The next proposition confirms the existence of TOS equilibria.

Proposition 4.1 (Existence of Time-of-Service Equilibrium) *Assume the capacity vector* $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$ *is given. There must exist at least one equilibrium arrival rate matrix* $\mathbf{\Lambda}^*$.

Showing the existence of an equilibrium Λ^* is equivalent to finding an arrival rate matrix Λ that satisfies (EC1) and (EC2) for all strategic customers. We use a technique developed by Beckmann et al. (1956) to verify the result. Let us consider the optimal solution to the following ancillary problem

$$\max_{\mathbf{\Lambda}} \quad \sum_{j=1}^{n}\sum_{i=1}^{m} \int_0^{\sum_{t=1}^{m}\Lambda_{tj}} (v_{ij} - c_{ij}W_j(x))\,dx \tag{4.3}$$

$$\text{s.t.} \quad \sum_{j=1}^{n} \Lambda_{ij} = \Lambda_i, i = 1, \ldots, m, \text{ and } \Lambda_{ij} \geq 0,$$

where $W_j(x) = \left(\mu_j - \Lambda^b_j - x\right)^{-1}$. Apparently, this problem has a convex feasible region. We can further verify that the objective function is concave in decision variable Λ_{ij}'s, $i = 1, \ldots, m$ and $j = 1, \ldots, n$. Therefore, the ancillary problem (4.3) is convex and admits at least one optimal solution. By invoking the Kuhn-Tucker conditions, we show that (EC1) and (EC2) are achieved at optimality. In other words, any optimal solution to the ancillary problem (4.3) corresponds to an equilibrium arrival rate matrix $\mathbf{\Lambda}^*$.

The validity of Proposition 4.1 does not rely on the assumption of an M/M/1 system. The TOS equilibrium exists for many other queueing systems, e.g., Erlang-C systems which are widely used in call center modeling.

We note that the equilibrium arrival matrix $\mathbf{\Lambda}^*$ may not be unique. However, the next result shows that the arrival rate vector is unique in all equilibria.

Proposition 4.2 (Uniqueness of Arrivals) *Assume the capacity vector* $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$ *is given. The total arrival rate of strategic customers to period* j, $j = 1, \ldots, n$, *is the same in all time-of-service equilibria. Alternatively, if* $\boldsymbol{\Lambda}^*$ *and* $\widehat{\boldsymbol{\Lambda}}^*$ *are two equilibrium arrival rate matrices, it must be that for any* $j = 1, \ldots, n$,

$$\sum_{i=1}^{m} \Lambda_{ij}^* = \sum_{i=1}^{m} \widehat{\Lambda}_{ij}^*.$$

Since the arrival rates of base customers are fixed, Proposition 4.2 also implies the total arrival rate λ_j, including both base and strategic customers, to period j is also unique. As a result, customers to the same time period would incur the same expected waiting time in all equilibria. Nonetheless, the compositions of strategic customers in each period can be different in distinct equilibria. Namely, if $\boldsymbol{\Lambda}^*$ and $\widehat{\boldsymbol{\Lambda}}^*$ are both TOS equilibria, there must exist a period j such that $\Lambda_{ij}^* \neq \widehat{\Lambda}_{ij}^*$ for a certain type i.

4.5 Capacity Allocation under TOS Equilbrium

Our discussion in the previous section presumes that service capacities in all periods are fixed. In this section, we first vary the capacity allocation and illustrate its impact on customer equilibrium TOS choices. We then account for these equilibrium choices and consider the capacity allocation problems. We propose two models to depict two different scenarios. In Section 4.5.3, we focus on the problem in which total capacity is fixed and difficult to increase in the short run. This model is of importance for many public service systems, in particular, emergency rooms, in which shortage of physician has been widely reported. Our model in this subsection addresses these service providers' challenges in capacity management. Section 4.5.4 adds a cost perspective and discusses optimal capacity decisions over time when capacities have to be purchased at time-varying rates. In both models, the service provider determines capacities designated to each period and strategic customers choose their TOS's in response. This endogeneity raises several difficulties. First, the underlying capacity allocation problems are subject to customer TOS equilibrium constraints. This type of optimization problems in general does not allow a tractable analytical solution and may be numerically challenging. We instead consider the so-called first-best (FB) problem which characterizes the system-wide optimality by allowing the provider to control both capacity decisions and customer routing. However, strategic customers determine their TOS selfishly, which results in two issues: (i) The FB capacity allocation might not be incentive compatible with customer TOS equilibria. (ii) System efficiency might be jeopardized due to strategic customers'

self-interested actions. We will address these issues as well in this section.

4.5.1 Impacts of Capacity Allocation on TOS Choices

Proposition 4.1 establishes the existence of TOS equilibria for a fixed capacity vector $\boldsymbol{\mu}$, but does not reveal how TOS equilibria vary as capacity allocation changes. We illustrate the effect of the capacity allocation on customer equilibrium TOS choices by an example.

A Two-Period Two-Type Example. Consider a two-period model with 2 types of strategic customers. When delays are not taken into account, strategic customers of types 1 and 2 prefer service in periods 1 and 2 respectively, i.e. $v_{11} > v_{12}$ and $v_{22} > v_{21}$. Unit delay costs for both types are time-invariant. In other words, $c_{11} = c_{12} = c_1$ and $c_{21} = c_{22} = c_2$. We fix the total capacity available to the two periods as a constant μ but vary the allocation continuously. Note that in this particular example, the TOS equilibrium is unique for all allocations.

Figure 4.1: Equilibrium of Two-Period Two-Type Case

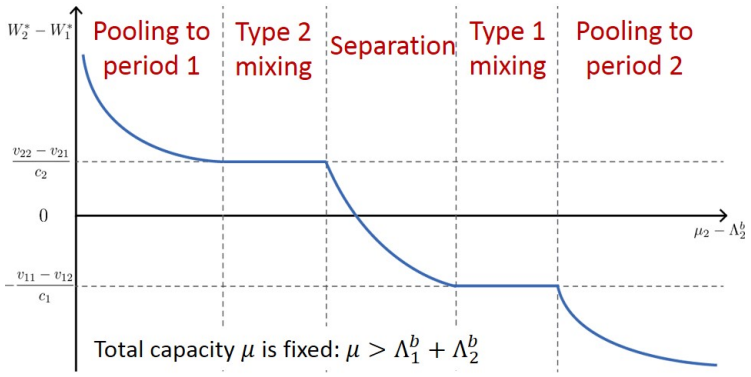


Figure 4.1 illustrates equilibrium TOS choices as a function of the excess capacity in period 2 after serving the base customers, i.e., $\mu_2 - \Lambda_2^b$. Let W_i^*, $i = 1, 2$, be the expected waiting time of period i in equilibrium. The vertical axis accordingly represents the expected delay difference $W_2^* - W_1^*$ in customer equilibrium TOS choices.

When the excess capacity $\mu_2 - \Lambda_2^b$ is barely above zero, congestion in period 2 is much severer than that in period 1. Although type 2 customers compromise and join a less favorable time period, their equilibrium net utility $v_{21} - c_2 W_1^*$ still surpasses what they would earn if they joined period 2, i.e., $v_{21} - c_2 W_1^* > v_{22} - c_2 W_2^*$. Thereby, all strategic customers are pooled to period 1 if $\mu_2 - \Lambda_2^b$ is marginally above zero. As the excess capacity $\mu_2 - \Lambda_2^b$ increases, congestion in period 2 is gradually relieved and the expected delay difference decreases. The reduced discrepancy incentivizes type 2 customers to switch

back to their preferred period. Nevertheless, when $\mu_2 - \Lambda_2^b$ is only at an intermediate level, accommodating only a fraction of type 2 customers would raise the delay difference to an equilibrium state in which $v_{21} - c_2 W_1^* = v_{22} - c_2 W_2^*$. In consequence, type 2 customers are mixed in two time periods in equilibrium until the excess capacity $\mu_2 - \Lambda_2^b$ reaches a higher level such that both types of strategic customers are separately served in their favorable periods. Successive escalation in period 2's excess capacity continues the decline in expected delay difference, which not only reinforces type 2 customers' willingness to stay but also starts to attract type 1 customers. The tendency of serving more customers in period 2 stops when the excess capacity $\mu_2 - \Lambda_2^b$ is sufficiently large and all type 1 customers eventually join period 2.

Figure 4.1 simply illustrates the effect of capacity allocation on equilibrium arrivals and we will account for this effect and deliberate how to optimally manage the capacity allocation over time in order to achieve maximum system efficiency.

4.5.2 System Welfare as a Performance Measure

We use the system welfare as a performance measure. For a given capacity allocation $\boldsymbol{\mu}$ and an arrival rate $\boldsymbol{\Lambda}$, the system welfare includes net utilities of base and strategic customers, i.e.,

$$S(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = \sum_{j=1}^{n} \left(v_j^b - c_j^b W_j(\mu_j, \lambda_j)\right) \Lambda_j^b + \sum_{j=1}^{n} \sum_{i=1}^{m} \left(v_{ij} - c_{ij} W_j(\mu_j, \lambda_j)\right) \Lambda_{ij} \quad (4.4)$$

$$= \sum_{j=1}^{n} v_j^b \Lambda_j^b + \sum_{j=1}^{n} \sum_{i=1}^{m} v_{ij} \Lambda_{ij} - \sum_{j=1}^{n} C_j(\lambda_j) W_j(\mu_j, \lambda_j), \quad (4.5)$$

where $C_j(\lambda_j)$ is defined in (4.2). We consider the system welfare as a performance measure for two reasons. First, the system welfare serves as a good benchmark to evaluate system performance as a whole. Second, this performance measure is plausible for a nonprofit facility whose first priority is to provide high quality service with high efficiency to the public. Instances of such facilities include emergency departments in Canada, United Kingdom and other countries or call centers for government services, which may operate under a large workforce with more than 500 employees (Pelleau et al. 2014). Furthermore, for many private firms, the main responsibility of their call centers is to retain customers and protect the organizations' greatest asset–their customers (Desmarais n.d.). From that perspective, it seems also reasonable to use total customer utility as a performance measure.

4.5.3 Capacity Allocation with Fixed Total Capacity

In this section, we restrict our attention to the case where the facility has to operate under a limited total capacity. This capacity constraint can be due to several reasons. On the one hand, managers of many service facilities, e.g., call centers and hospitals, receive fixed amounts of operating budget from finance departments of corporations or governments, which limits the amount of available capacity. On the other hand, capacity can also be a resource with inadequate supply under certain circumstances, particularly in health care industry. The shortage of doctors and nurses has been well reported and documented by media (cf. The Globe and Mail 2013) or independent heath care organizations (cf. Council on Physician and Nurse Supply 2007). Therefore, efficient operation strategy may be of particular interest in the fixed total capacity situation.

4.5.3.1 No-Balking Case

We first consider the problem without balking. This model suggests useful insights into capacity management, customer routing and incentive alignment. We will relax this assumption and explore its impact in Section 4.5.3.2.

The provider determines his intertemproal capacity allocation $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$ of a fixed total amount μ, which can be interpreted as a projection of the total number of shifts the provider can assign. In response, strategic customers selfishly choose their TOS which results in the equilibrium arrival $\boldsymbol{\Lambda}^*(\boldsymbol{\mu})$. The provider's objective is to maximize the system welfare $S(\boldsymbol{\mu}, \boldsymbol{\Lambda}^*(\boldsymbol{\mu}))$ while accounting for customer responses in timing their arrivals. Mathematically, the provider solves the following capacity allocation problem with fixed resource, referred to as the CAFR problem,

$$
\begin{aligned}
(\mathbf{CAFR}) \quad \max_{\boldsymbol{\mu}} \quad & S(\boldsymbol{\mu}, \boldsymbol{\Lambda}^*(\boldsymbol{\mu})) && && \\
\text{s.t.} \quad & \mu_1 + \cdots + \mu_n = \mu, && && (4.6)\\
& \textstyle\sum_{j=1}^{n} \Lambda_{ij}^*(\boldsymbol{\mu}) = \Lambda_i,\ \Lambda_{ij}^*(\boldsymbol{\mu}) \geq 0, && i = 1, \ldots, m,\ j = 1, \ldots, n && (4.7)\\
& \lambda_j(\boldsymbol{\Lambda}^*(\boldsymbol{\mu})) = \Lambda_j^b + \textstyle\sum_{i=1}^{m} \Lambda_{ij}^*(\boldsymbol{\mu}) < \mu_j, && j = 1, \ldots, n && (4.8)\\
& U_{ik}(\boldsymbol{\Lambda}^*(\boldsymbol{\mu})) = U_{ik'}(\boldsymbol{\Lambda}^*(\boldsymbol{\mu})), && \forall\ k, k' \in \mathcal{P}_i(\boldsymbol{\mu}) && (4.9)\\
& U_{ik}(\boldsymbol{\Lambda}^*(\boldsymbol{\mu})) \geq U_{ik'}(\boldsymbol{\Lambda}^*(\boldsymbol{\mu})), && \forall\ k \in \mathcal{P}_i(\boldsymbol{\mu}) \text{ and } k' \in \mathcal{Z}_i(\boldsymbol{\mu}) && (4.10)
\end{aligned}
$$

where $\mathcal{P}_i(\boldsymbol{\mu}) = \{j \mid \Lambda_{ij}^*(\boldsymbol{\mu}) > 0\}$ and $\mathcal{Z}_i(\boldsymbol{\mu}) = \{j \mid \Lambda_{ij}^*(\boldsymbol{\mu}) = 0\}$. Although the provider could choose not to exhaust all capacity, it will never be optimal to do so. The system welfare can always be improved by allocating any residual capacity to any period in order to alleviate congestion. Thus, we impose the binding capacity constraint $\sum_j \mu_j = \mu$ without loss of generality. In the CAFR problem, we allow strategic customers to self-

select TOS, yet they are not permitted to balk. In other words, every customer will be eventually served by the provider as shown by the demand constraint $\sum_j \Lambda^*_{ij}(\boldsymbol{\mu}) = \Lambda_i$. We will relax this condition in Section 4.5.3.2. In order to achieve a stable system, allocated capacity to each period must exceed the demand. Therefore, we have the stability constraint (4.8). Finally, strategic customers choose their TOS's and the system reaches an equilibrium when (4.9) and (4.10) are simultaneously satisfied.

The CAFR problem belongs to a special class of optimization problems, called Mathematical Programming with Equilibrium Constraints (MPEC). Directly solving the MPEC problems is challenging since the feasible region is ill-conditioned due to the equilibrium constraints (Luo et al. 1996). For this reason, we first consider a relatively simpler problem, the so-called first-best (FB) problem

$$(\mathbf{CAFR_{fb}}) \quad \max_{\boldsymbol{\mu},\boldsymbol{\Lambda}} \; S(\boldsymbol{\mu},\boldsymbol{\Lambda}) \text{ s.t. } \sum_{j=1}^{n} \mu_j = \mu,\; \sum_{j=1}^{n} \Lambda_{ij} = \Lambda_i,\; \lambda_j = \Lambda^b_j + \sum_{i=1}^{m} \Lambda_{ij} < \mu_j, \text{ and } \Lambda_{ij} \geq 0$$

for $j = 1, \ldots, n$ and $i = 1, \ldots, m$.

In the above FB problem of CAFR, which we refer to as CAFR_{fb}, the provider not only assigns capacity to each period but also controls customer routing decisions. Hence, the TOS equilibrium constraint in the CAFR problem becomes irrelevant. Unfortunately, the CAFR_{fb} problem is still difficult to solve because the objective function is lack of concavity or even quasi-concavity. Rather than deriving the full solution of CAFR_{fb}, we characterize two important properties of its optimal solution.

First, we note from (4.5) that the capacity allocation only affects the total system delay cost but not the overall rewards customers collect. Therefore, capacity allocation is an independent decision when the customer routing $\boldsymbol{\Lambda}$ is given. We thus analytically characterize the optimal capacity allocation rule for given arrivals as shown in next proposition.

Proposition 4.3 (Optimal Capacity Allocation) *Assume the arrival rate matrix* $\boldsymbol{\Lambda}$ *is given and let* $C_j(\lambda_j)$*, defined in* (4.2)*, be the corresponding unit delay cost in period* j*,* $j = 1, \ldots, n$*. The following waiting cost minimization problem*

$$\min_{\boldsymbol{\mu} > \boldsymbol{\lambda}} \quad \sum_{j=1}^{n} C_j(\lambda_j) \cdot W_j(\mu_j \mid \lambda_j) \text{ s.t. } \sum_{j=1}^{n} \mu_j = \mu \tag{4.11}$$

is convex and has a unique optimal solution

$$\mu^*_j = \lambda_j + \frac{\sqrt{C_j(\lambda_j)}}{\sum_{k=1}^{n} \sqrt{C_k(\lambda_k)}} \left(\mu - \sum_{k=1}^{n} \lambda_k \right). \tag{4.12}$$

The optimal capacity formula (4.12) essentially specifies the best way to allocate the excess capacity after the system stability has been established. The provider divides the excess capacity $\mu - \sum_k \lambda_k$ in a way that is proportional to the square root of the waiting cost of each period.

The optimal capacity (4.12) implies that customer arrivals to period j affect not only period j's delay but also others'. This effect results from the fact that all periods share the same available capacity. An increase in period j's customer volume leads to a higher capacity allocation to period j, which lowers available capacity to other periods and increases delays of other time periods.

Note that the delay cost $C_j(\lambda_j)W_j(\mu_j \mid \lambda_j)$ in each period is decreasing and convex w.r.t. the capacity level μ_j. When allocating any additional capacity, the provider would prefer using it to a period with the largest marginal reduction in the delay cost. As the capacity assigned to that period increases, the marginal delay cost reduction decreases due to the convexity. Then, whenever an alternative period offers a better marginal reduction, the provider starts to invest capacity in the alternative period. As a result, at the optimal allocation, it must be that the marginal reductions in delay cost of all periods are the same. Mathematically, it means

$$C_1(\lambda_1)\frac{\partial W_1(\mu_1^* \mid \lambda_1)}{\partial \mu_1} = C_1(\lambda_2)\frac{\partial W_2(\mu_2^* \mid \lambda_2)}{\partial \mu_2} = \cdots = C_n(\lambda_n)\frac{\partial W_n(\mu_n^* \mid \lambda_n)}{\partial \mu_n}. \tag{4.13}$$

In addition to describing the optimal capacity allocation for a given arrival $\mathbf{\Lambda}$, (4.13) also plays an important role in achieving incentive compatibility with customer selfish TOS choices, which we will discuss later.

Proposition 4.3 also offers a way to transform the CAFR_{fb} problem, a simultaneous capacity allocation and customer routing problem, into a pure customer routing problem. Apply (4.12) to the system welfare $S(\boldsymbol{\mu}, \mathbf{\Lambda})$ in (4.5). We have the system welfare as a function of customer routing decisions $\mathbf{\Lambda}$ only,

$$S(\boldsymbol{\mu}^*, \mathbf{\Lambda}) = \sum_{j=1}^{n} v_j^b \Lambda_j^b + \sum_{j=1}^{n}\sum_{i=1}^{m} v_{ij}\Lambda_{ij} - \sum_{j=1}^{n} C_j(\lambda_j)W_j(\mu_j^* \mid \lambda_j) \tag{4.14}$$

$$= \sum_{j=1}^{n} v_j^b \Lambda_j^b + \sum_{j=1}^{n}\sum_{i=1}^{m} v_{ij}\Lambda_{ij} - \frac{\left(\sum_{j=1}^{n}\sqrt{c_j^b\Lambda_j^b + \sum_{i=1}^{m} c_{ij}\Lambda_{ij}}\right)^2}{\mu - \sum_{j=1}^{n}\Lambda_j^b - \sum_{i=1}^{m}\Lambda_i}, \tag{4.15}$$

where $\boldsymbol{\mu}^* = \boldsymbol{\mu}^*(\mathbf{\Lambda})$ is the optimal capacity vector obtained from (4.12).[4] Therefore,

[4]For notational parsimony, we write the optimal capacity allocation without arguments, with the understanding that $\boldsymbol{\mu}^*$ is always a function of the given arrival rate matrix $\mathbf{\Lambda}$.

solving the CAFR_{fb} problem is equivalent to finding the solution of

$$\max_{\mathbf{\Lambda}} \; S(\boldsymbol{\mu}^*, \mathbf{\Lambda}) \text{ s.t. } \sum_{j=1}^{n} \Lambda_{ij} = \Lambda_i, \; \lambda_j = \Lambda_j^b + \sum_{i=1}^{m} \Lambda_{ij}, \text{ and } \Lambda_{ij} \geq 0$$

for $j = 1, \ldots, n$ and $i = 1, \ldots, m$. Acquiring a closed-form solution to the optimal $\mathbf{\Lambda}^*$ is still difficult. Nonetheless, we notice that the optimal customer routing decision must satisfy the "no-mixing" property as the next proposition states.

Proposition 4.4 (Optimal Customer Routing) *Let* $\mathbf{\Lambda}^* = \arg\max_{\mathbf{\Lambda}} S(\boldsymbol{\mu}^*, \mathbf{\Lambda})$, *where* $\boldsymbol{\mu}^* = \boldsymbol{\mu}^*(\mathbf{\Lambda})$ *is the optimal capacity allocation in* (4.12) *for a given* $\mathbf{\Lambda}$. *Then, for each row* $\mathbf{\Lambda}_i^* = (\Lambda_{i1}, \ldots, \Lambda_{in})$, $i = 1, \ldots, m$, *only one entry* Λ_{ij}^* *can be strictly positive and it equals* Λ_i. *In other words, all strategic customers of the same type would be served in exactly one time period at optimality.*

The above result shows that under *endogenous* capacity, the provider finds it optimal to serve all strategic customers in the same time period. Yet, it could well be optimal to serve multiple types in one period. Recall the two-period two-type example we discussed in Figure 4.1. In equilibrium, strategic customers of the same type may join two different periods. Proposition 4.4 clearly indicates allowing the same type customers to join multiple periods would not lead to the maximum efficiency. The provider tends to assigning strategic customers to either a period that offers higher service valuation or one that can provide less congestion. In the example of Figure 4.1, the provider would prefer the "separation" strategy if both types are strongly in favor of particular periods, whereas pooling both types in one period may sustain as a superior strategy when one type is relatively TOS insensitive.

Although the CAFR_{fb} problem provides guidance on how to achieve system-wide optimality via capacity and customer allocation, the provider may not be able to control customer TOS choices. It is strategic customers themselves who determine their arrival times that maximize their expected utilities. Proposition 4.5 shows that the first-best solution will also arise under strategic customer self-interested TOS choices.

Proposition 4.5 (Incentive Compatibility of the CAFB Problem) *The optimal solution to the* CAFR_{fb} *problem also satisfies the TOS equilibrium constraint in the CAFR problem, i.e., the first-best solution to the CAFR problem is also incentive compatible with strategic customer self-interest TOS choices.*

Since the FB solution to the CAFR problem is incentive compatible, strategic customer self-interested TOS choices are guaranteed to induce a system-wide optimal arrival

rate pattern. In contrast to the previous literature on selfish routing in computer and communication network (Roughgarden 2005), this result highlights the advantage of being able to adjust capacity levels to cope with self-interested decisions. It is well-known that routing decisions of self-interested users through a congested network, which has no central authority, in general cause performance inefficiency (Roughgarden 2005). Such inefficiency is usually referred to as the price of anarchy (PoA), which quantifies the negative impact of selfish action relative to system optimality. It has been shown that the PoA can be arbitrarily high in many settings (Roughgarden and Tardos 2002, Friedman 2004). In contrast, Proposition 4.5 demonstrates the advantage of capacity allocation in achieving system-wide optimality even under customer selfish choices.

Propositions 4.4 and 4.5 together also imply that customers use pure strategies when choosing TOS under endogenous capacity decisions. This result is in sharp contrast to the result from stochastic location models with congestion, where capacity is fixed. As illustrated in Berman and Krass (2015) (Example 1), customers in general use mixed strategies in choosing visiting facilities, i.e., they randomize among several facilities. When restricting customer choices to pure strategies, the system may not be able to reach an equilibrium at all depending on the given capacity at each server. This observation also coincides with our demonstrating example in Figure 4.1. The reason that our model can induce pure strategies is because service rates are decision variables in our model, rather than given parameters. Whenever a mixed strategy is employed and the system loses its optimality in efficiency, the provider can re-allocate the capacity to divert customers to the servers or time periods that are optimal for the system.

We next discuss the intuition for Proposition 4.5. Suppose it is optimal for the provider to serve all type i customers in period k. The following must hold for any other period $k' \neq k$. At optimality, where all type i customers are served in period k and none in any other period k' by Proposition 4.4, it must be that the marginal system welfare w.r.t. an increase in the type i arrival rate to period k exceeds the marginal system welfare w.r.t. an increase in the type i arrival rate to period k', i.e.,

$$v_{ik} - c_{ik} W_k(\mu_k^*, \lambda_k^*) - C_k(\lambda_k^*) \frac{\partial W_k(\mu_k^*)}{\partial \Lambda_{ik}} \Big|_{\Lambda_{ik}=\Lambda_i} \geq v_{ik'} - c_{ik'} W_{k'}(\mu_{k'}^*, \lambda_{k'}^*) - C_{k'}(\lambda_{k'}^*) \frac{\partial W_{k'}(\mu_k^*)}{\partial \Lambda_{ik'}} \Big|_{\Lambda_{ik'}=0}. \tag{4.16}$$

In (4.16) the marginal system welfare is written as the difference between the individual utility of the additional customer flow minus the externality inflicted on the rest of the system. Since we consider an M/M/1 system, marginal effect on expected delay of serving more customers is equivalent to that of reducing the capacity by the same amount, i.e.,

$\frac{\partial W_j}{\partial \Lambda_{ij}} = -\frac{\partial W_j}{\partial \mu_j}$. By (4.16), we have

$$v_{ik}-c_{ik}W_k(\mu_k^*,\lambda_k^*)+C_k(\lambda_k^*)\frac{\partial W_k(\mu_k^*)}{\partial \mu_k}\Big|_{\Lambda_{ik}=\Lambda_i} \geq v_{ik'}-c_{ik'}W_{k'}(\mu_{k'}^*,\lambda_{k'}^*)+C_{k'}(\lambda_{k'}^*)\frac{\partial W_{k'}(\mu_{k'}^*)}{\partial \mu_{k'}}\Big|_{\Lambda_{ik'}=0}. \tag{4.17}$$

Moreover, as shown in (4.13), the marginal reductions in delay cost of all periods are the same under the optimal capacity allocation. Therefore, (4.17) implies

$$v_{ik} - c_{ik}W_k(\mu_k^*,\lambda_k^*) \geq v_{ik'} - c_{ik'}W_{k'}(\mu_{k'}^*,\lambda_{k'}^*), \tag{4.18}$$

which indicates that type i customers prefer period k to other alternatives.

The result in Proposition 4.5 hinges on the assumption of exponential service times; however, the results of Dewan and Mendelson (1990) suggest that the efficiency loss may be small when this assumption is relaxed. Moreover, the flexibility to adjust capacity allocation is essential in aligning the provider's and customers' incentives. It has been shown that self-interested routing inevitably causes system performance loss when each server works in an M/M/1 manner with fixed capacity (e.g., Friedman 2004, Haviv and Roughgarden 2007).

In summary, when balking is not allowed or the provider has to serve all customers, the optimal solution to the CAFR problem coincides with the system FB solution. At optimality, strategic customers of the same type are served in the same period under the equilibrium TOS choices.

4.5.3.2 Balking Case

We now relax the assumption that balking is not allowed and also assume that the provider controls the admission rates of different customer segments. Let $\overline{\Lambda}_j^b$ and $\overline{\Lambda}_i$ be the maximum potential arrival rates for base customers of period j and type i strategic customers, respectively. We consider the following capacity allocation problem under fixed total capacity with balking,

$$\begin{aligned}
(\mathbf{CAFR^b}) \quad \max_{\boldsymbol{\mu}} \quad & S(\boldsymbol{\mu}, \boldsymbol{\Lambda}^*(\boldsymbol{\mu})) & & \\
\text{s.t.} \quad & \mu_1 + \cdots + \mu_n = \mu, & & (4.19)\\
& 0 \leq \Lambda_j^b \leq \overline{\Lambda}_j^b,\ U_j^b(\boldsymbol{\Lambda}^*(\boldsymbol{\mu})) \geq 0, & j = 1,\ldots,n & (4.20)\\
& \textstyle\sum_{j=1}^n \Lambda_{ij}^*(\boldsymbol{\mu}) = \Lambda_i < \overline{\Lambda}_i,\ \Lambda_{ij}^*(\boldsymbol{\mu}) \geq 0, & i = 1,\ldots,m,\ j = 1,\ldots,n & (4.21)\\
& \lambda_j(\boldsymbol{\Lambda}^*(\boldsymbol{\mu})) = \Lambda_j^b + \textstyle\sum_{i=1}^m \Lambda_{ij}^*(\boldsymbol{\mu}) < \mu_j, & j = 1,\ldots,n & (4.22)\\
& U_{ik}(\boldsymbol{\Lambda}^*(\boldsymbol{\mu})) = U_{ik'}(\boldsymbol{\Lambda}^*(\boldsymbol{\mu})) \geq 0, & \forall\ k,k' \in \mathcal{P}_i(\boldsymbol{\mu}) & (4.23)\\
& U_{ik}(\boldsymbol{\Lambda}^*(\boldsymbol{\mu})) \geq U_{ik'}(\boldsymbol{\Lambda}^*(\boldsymbol{\mu})), & \forall\ k \in \mathcal{P}_i(\boldsymbol{\mu}) \text{ and } k' \in \mathcal{Z}_i(\boldsymbol{\mu}) & (4.24)
\end{aligned}$$

where $\mathcal{P}_i(\boldsymbol{\mu}) = \{j \mid \Lambda^*_{ij}(\boldsymbol{\mu}) > 0\}$ and $\mathcal{Z}_i(\boldsymbol{\mu}) = \{j \mid \Lambda^*_{ij}(\boldsymbol{\mu}) = 0\}$. The CAFR$^{\text{b}}$ problem is similar to the CAFR problem. The service provider allocates the total available capacity μ accounting for strategic customers' equilibrium responses to the potential differences in delays over time. However, the provider now does not need to serve every customer as specified in constraints (4.20) and (4.21). Moreover, constraints (4.21) and (4.23) require all base and strategic customers receiving non-negative utilities upon service completion. Otherwise, customers can choose to balk – an option yielding zero reward.

Analogous to the no-balking case, we first ignore the TOS equilibrium constraints and further the non-negative net utility constraints. Let $\boldsymbol{\Lambda}^b = (\Lambda^b_1, \ldots, \Lambda^b_n)$ and consider the corresponding FB problem

$$(\mathbf{CAFR^b_{fb}}) \max_{\boldsymbol{\mu},\boldsymbol{\Lambda},\boldsymbol{\Lambda}^b} S(\boldsymbol{\mu}, \boldsymbol{\Lambda}) \text{ s.t. } \sum_{j=1}^{n} \mu_j = \mu,\ \sum_{j=1}^{n} \Lambda_{ij} \le \overline{\Lambda}_i,\ \lambda_j = \Lambda^b_j + \sum_{i=1}^{m} \Lambda_{ij} < \mu_j,\ \Lambda_{ij} \ge 0 \text{ and } 0 \le \Lambda^b_j \le \overline{\Lambda}^b_j$$

for $j = 1, \ldots, n$ and $i = 1, \ldots, m$, in which the provider is able to control the capacity allocation, admission rate of every customer segment, and routing decisions.

Denote an optimal solution to the CAFR$^{\text{b}}_{\text{fb}}$ by $(\boldsymbol{\mu}^*, \boldsymbol{\Lambda}^*, \boldsymbol{\Lambda}^{b*})$ and the optimal arrival rate of type i strategic customers $\Lambda^*_i = \sum_{j=1}^{m} \Lambda^*_{ij}$. Note that the "no-mixing" customer routing property in Propositions 4.4 holds for any given $\boldsymbol{\Lambda}$, in particular, when $\boldsymbol{\Lambda} = \boldsymbol{\Lambda}^*$, and the optimal capacity allocation rule in Proposition 4.3 applies to any known $\boldsymbol{\Lambda}$ and $\boldsymbol{\Lambda}^b$, in particular, when $\boldsymbol{\Lambda} = \boldsymbol{\Lambda}^*$ and $\boldsymbol{\Lambda}^b = \boldsymbol{\Lambda}^{b*}$. Hence, the incentive compatibility with customer self-interested TOS choices is also achieved by the FB solution $(\boldsymbol{\mu}^*, \boldsymbol{\Lambda}^*, \boldsymbol{\Lambda}^{b*})$ to the CAFR$^{\text{b}}_{\text{fb}}$ problem.

Given that all the results in the no-balking case, Propositions 4.3–4.5, still hold at the optimal arrival rates $\boldsymbol{\Lambda}^{b*}$ and $\boldsymbol{\Lambda}^{b*}$, the service provider only needs to concern how to induce the optimal arrival rates $\boldsymbol{\Lambda}^*$ and $\boldsymbol{\Lambda}^{b*}$. It is well known that pricing can be used as a leverage to manage customer flows (e.g., Naor 1969, Mendelson and Whang 1990). However, in our setting, there are multiple types of customers with different delay sensitivities and it is possible that several types might be simultaneously served in the same time period. Nevertheless, Proposition 4.6 shows that a single time-invariant and type-independent price is sufficient to induce the optimal rates of all types.

Proposition 4.6 (Pricing for Optimal Arrivals) *Assume that $\boldsymbol{\Lambda}^*$ and $\boldsymbol{\Lambda}^{b*}$ are the optimal arrival rates for strategic and base customers and $\lambda^*_j = \Lambda^{b*}_j + \sum_{i=1}^{m} \Lambda^*_{ij}$, $j = 1, \ldots, n$, is the optimal total arrival rate to period j, $j = 1, \ldots, n$. The provider can charge a time-invariant and type-independent price*

$$p = \left(\sum_{j=1}^{n} \sqrt{C_j(\lambda^*_j)} \right)^2 \left(\mu - \sum_{j=1}^{n} \lambda^*_j \right)^{-2} \tag{4.25}$$

to both strategic and base customers in order to induce the optimal total arrival rate to each period.

We use strategic customers to discuss this result. The same rationale applies to base customers as well. Consider the marginal value of serving type i customers when capacity is optimally allocated according to (4.12). As mentioned before, strategic customers of the same type would be served in the same time period, say period j^*. From (4.14), the marginal value of serving type i customers is

$$\frac{\partial S^*}{\partial \Lambda_i} = \frac{\partial S^*}{\partial \Lambda_{ij^*}} = v_{ij^*} - c_{ij} W^*_{j^*} - C_{j^*}(\Lambda_i)\frac{\partial W^*_{j^*}}{\partial \Lambda_i} - \sum_{j \neq j^*} C_j \frac{\partial W^*_j}{\partial \Lambda_i},$$

where $W^*_j(\Lambda_i) = W_j(\mu^*_j(\Lambda_i) \mid \lambda_j)$, $j = 1, \ldots, n$, and $C_{j^*}(\Lambda_i) = C_{j^*}(\lambda_{j^*}(\Lambda_i))$ are all functions of Λ_i and S^* is a shorthand of $S(\boldsymbol{\mu}^*, \boldsymbol{\Lambda})$ in (4.14). The first two terms $v_{ij^*} - c_{ij^*} W^*_{j^*}$ represent the individual net utility and the last two terms capture the externality. Since customer self-interested decisions ignore the externality they inflicts on the system, the provider has to charge type i customers a price that is equal to the externality cost, which is characterized by (4.25), to align their incentives in order to induce the optimal arrival rate. Moreover, the provider's optimal capacity allocation policy balances externalities inflicted on others such that irrespective of which types are served and in which period they are served, externalities are always the same. This property eventually allows a uniform price to achieve the optimal arrivals for all customers who may choose to join different time periods for services.

Finally, at the optimal solution $(\boldsymbol{\mu}^*, \boldsymbol{\Lambda}^*, \boldsymbol{\Lambda}^{b*})$, it must be that $\partial S^*/\partial \Lambda_i \geq 0$ for $i = 1, \ldots, m$, which implies net utilities that type i customers' net utilities $v_{ij^*} - c_{ij^*} W^*_{j^*}$ are non-negative at optimality.

In summary, when allowing customers to balk and the provider to control arrival rates, the service provider only needs to consider a mechanism to induce the optimal arrivals. A uniform price that equals the externality cost at optimal arrivals $\boldsymbol{\Lambda}^*$ and $\boldsymbol{\Lambda}^{b*}$ is sufficient for that purpose.

4.5.4 Capacity Allocation with Time-Varying Capacity Cost

The previous model may well depict a situation in which the total capacity is inadequate, such as in health care industry. In this section, we consider another model that attempts to balance the expected benefits of increasing capacity against the capacity cost in the same vein as the long-run model in Mendelson (1985). Specifically, capacity in period j, $j = 1, \ldots, n$, can be purchased at a uint cost b_j and is never in short supply. We allow

time-dependent capacity cost, which reflects a natural industrial reality. In traditional call centers, outsourcing part of the service offshare is a common practice to reduce unit capacity cost. Depending on the the ratio of onshare and offshare agents at different times, unit capacity costs vary over time. In the instance of recent innovative work-from-home call centers (e.g., LiveOps and Arise Virtual Solutions), providers have to offer a higher hourly salary in high volume time to attract enough workers (Gurvich et al. 2014). As a result, providers' unit capacity costs would fluctuate over time.

We again use system welfare as a performance measure. However, the provider has to deduct the capacity cost from the total welfare. Therefore, we denote the system net welfare as

$$NS(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = S(\boldsymbol{\mu}, \boldsymbol{\Lambda}) - \sum_{j=1}^{n} b_j \mu_j,$$

where $S(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ is defined in (4.5) and the second term represents the provider's overall capacity cost for a chosen capacity vector $\boldsymbol{\mu}$. This performance measure considers both customer delays and the provider capacity investment cost.

For simplicity, we again first consider the no-balking case and later discuss how our results would change if balking is allowed and the provider can control the admission rates of all customer segments.

4.5.4.1 No-balking Case

The provider chooses his intertemporal capacity plan $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$ to maximize the system net welfare. We assume that capacity for each period is always available at a unit cost b_j. This assumption distinguishes the CAFR problem with the current capacity allocation (CA) problem. Hence, the capacity constraint is irrelevant and we express the provider's problem as

$$(\mathbf{CA}) \quad \max_{\boldsymbol{\mu}} \; NS(\boldsymbol{\mu}, \boldsymbol{\Lambda}) \;\text{ s.t. }\; (4.7),\; (4.8),\; (4.9),\; \text{and } (4.10),$$

where (4.7)-(4.10) are the same constraints as in the CAFR problem. Due to the equilibrium constraints (4.9) and (4.10), it is challenging to directly solve the CA problem. We thus drop the customer TOS equilibrium constraints and consider the corresponding FB problem

$$(\mathbf{CA_{fb}}) \quad \max_{\boldsymbol{\mu}, \boldsymbol{\Lambda}} \; NS(\boldsymbol{\mu}, \boldsymbol{\Lambda}) \text{ s.t. } \quad \sum_{j=1}^{n} \Lambda_{ij} = \Lambda_i,\; \lambda_j = \Lambda_j^b + \sum_{i=1}^{m} \Lambda_{ij} < \mu_j,\; \text{and } \Lambda_{ij} \geq 0$$

for $j = 1, \ldots, n$ and $i = 1, \ldots, m$, in which the provider is assumed to be able to control customer routing in addition to the capacity allocation. As in the model considered above, it is also difficult to obtain a full solution to the CA problem. We characterize the optimal capacity levels over time for given intertemporal arrival rates and establish the "no-mixing" property of the customer routing decisions.

Proposition 4.7 (Optimal Capacity Allocation) *Assume the arrival rate matrix* $\mathbf{\Lambda}$ *is given and let* $C_j(\lambda_j)$*, defined in* (4.2)*, be the corresponding unit delay cost in period* j*,* $j = 1, \ldots, n$*. The following cost minimization problem*

$$\min_{\boldsymbol{\mu} > \boldsymbol{\lambda} \geq 0} \quad \sum_{j=1}^{n} C_j(\lambda_j) \cdot W_j(\mu_j \mid \lambda_j) + \sum_{j=1}^{n} b_j \mu_j$$

is convex and has a unique optimal solution

$$\mu_j^* = \lambda_j + \sqrt{C_j(\lambda_j)/b_j}. \tag{4.26}$$

When the arrival rates are given, the capacity decisions reflect the economical tradeoff between total delay cost and the capacity investment. Moreover, this tradeoff arises in each time period independently for a given $\mathbf{\Lambda}$. That is, the capacity decision of a single period does not have any influence to that of another period, which differs from the CAFR problem. In essence, the provider stops his investment in capacity of each period if and only if the marginal waiting cost reduction of the focal period equals the marginal capacity cost increment, i.e.,

$$-C_j(\lambda_j) \frac{\partial W_j(\mu_j^* \mid \lambda_j)}{\partial \mu_j} = b_j, \ j = 1, \ldots, n. \tag{4.27}$$

We may interpret the optimal capacity condition (4.27) as a generalization of (4.13). In the CAFR_{fb} problem, each unit of the capacity is equally costly to the provider. Therefore, the provider does not need to account for differences in unit capacity costs when allocating the capacity and only needs to concern the delay cost. The fact that marginal waiting cost reductions in all periods are equalized at optimality actually reflects the indistinctive unit capacity cost. In contrast, if unit capacity costs are different over time, we obtain a general optimal condition (4.27).

The capacity allocation criterion (4.27) also has a critical implication in aligning customer incentives with the socially optimal FB solution. We will further discuss it later in this section.

The optimal capacity decision (4.26) allows us to reduce the CA_{fb} problem to a pure routing problem. Let $\boldsymbol{\mu}^* = \boldsymbol{\mu}^*(\boldsymbol{\Lambda})$ be the optimal capacity vector from (4.26). We write

$$\begin{aligned} NS(\boldsymbol{\mu}^*, \boldsymbol{\Lambda}) &= S(\boldsymbol{\mu}^*, \boldsymbol{\Lambda}) - \sum_{j=1}^{n} b_j \mu_j^* \\ &= \sum_{j=1}^{n} v_j^b \Lambda_j^b + \sum_{j=1}^{n}\sum_{i=1}^{m} v_{ij}\Lambda_{ij} - \sum_{j=1}^{n} C_j(\lambda_j) W_j(\mu_j^* \mid \lambda_j) - \sum_{j=1}^{n} b_j \mu_j^*. \\ &= \sum_{j=1}^{n} v_j^b \Lambda_j^b + \sum_{j=1}^{n}\sum_{i=1}^{m} v_{ij}\Lambda_{ij} - \sum_{j=1}^{n} C_j(\lambda_j)\sqrt{b_j/C_j(\lambda_j)} - \sum_{j=1}^{n} b_j \left(\lambda_j + \sqrt{C_j(\lambda_j)/b_j}\right). \end{aligned}$$

By (4.1) and (4.2),

$$\begin{aligned} NS(\boldsymbol{\mu}^*, \boldsymbol{\Lambda}) &= \sum_{j=1}^{n} v_j^b \Lambda_j^b - \sum_{j=1}^{n} b_j \Lambda_j^b + \sum_{j=1}^{n}\sum_{i=1}^{m} v_{ij}\Lambda_{ij} - \sum_{j=1}^{n}\left(b_j \sum_{i=1}^{m} \Lambda_{ij}\right) \\ &\quad -2\sum_{j=1}^{n}\sqrt{b_j\left(c_j^b \Lambda_j^b + \sum_{i=1}^{m} c_{ij}\Lambda_{ij}\right)}. \end{aligned} \tag{4.28}$$

Therefore, solving the CA_{fb} problem is equivalent to finding the optimal value of

$$(\mathbf{CA_{fb}}) \quad \max_{\boldsymbol{\Lambda}} \; NS(\boldsymbol{\mu}^*, \boldsymbol{\Lambda}) \text{ s.t. } \sum_{j=1}^{n} \Lambda_{ij} = \Lambda_i, \; \lambda_j = \Lambda_j^b + \sum_{i=1}^{m} \Lambda_{ij} < \mu_j, \text{ and } \Lambda_{ij} \geq 0.$$

Analogous to the fixed total capacity case, the optimal customer routing satisfies the "no-mixing" property. That is, it is optimal to serve all customers of a given type in the same time period.

Proposition 4.8 (Optimal Customer Routing) *Let* $\boldsymbol{\Lambda}^* = \arg\max_{\boldsymbol{\Lambda}} S(\boldsymbol{\mu}^*, \boldsymbol{\Lambda})$*, where* $\boldsymbol{\mu}^* = \boldsymbol{\mu}^*(\boldsymbol{\Lambda})$ *is the optimal capacity allocation in* (4.26) *for a given* $\boldsymbol{\Lambda}$*. Then, for each row* $\boldsymbol{\Lambda}_i^* = (\Lambda_{i1}, \ldots, \Lambda_{in})$*,* $i = 1, \ldots, m$*, only one entry* Λ_{ij}^* *can be strictly positive and it equals* Λ_i*. In other words, all strategic customers of the same type would be served in exactly one time period at optimality.*

Again, this result simply shows that potential shorter delay from pooling and service in a more favorable period are mutually exclusive. Any attempt to engross both would impair the net welfare.

Although the FB solution suggests the best way to operate the system, the provider may not be able to control customer routing. Given the allocated capacities over time and their attributes, customers make TOS decisions on their own. Unlike in the capacity

allocation model studied above, here the FB solution to the CA_{fb} is not necessarily compatible with customer self-interested choices. The provider must impose a pricing scheme to reconcile customer incentives with his.

Proposition 4.9 (Incentive Compatibility of the CA Problem) *The optimal solution to the CA_{fb} problem is incentive compatible with customer selfish TOS choices if the provider charges such that price differences are equal to the corresponding differences in unit capacity cost, i.e., $p_j - p_{j'} = b_j - b_{j'}$ for any $j, j' = 1, \ldots, n$, where p_j is the service price for period j.*

The proposition above indicates that the incentive compatibility can be achieved as long as price differences in TOS equal the corresponding capacity cost differences. The intuition for this result is similar to the one in the fixed capacity model, but with a minor variation.

By Proposition 4.8, the provider prefers serving all type i customers in one period, say period k, at optimality. This implies that the marginal net welfare, which equals additional individual utility minus the externality, w.r.t. an increase in the type i arrival rate to period k exceeds the marginal net welfare w.r.t. an increase in the type i arrival rate to any other period k', i.e.,

$$v_{ik} - c_{ik} W_k(\mu_k^*, \lambda_k^*) - C_k(\lambda_k^*) \frac{\partial W_k(\mu_k^*)}{\partial \Lambda_{ik}} \Big|_{\Lambda_{ik}=\Lambda_i} \geq v_{ik'} - c_{ik'} W_{k'}(\mu_{k'}^*, \lambda_{k'}^*) - C_{k'}(\lambda_{k'}^*) \frac{\partial W_{k'}(\mu_k^*)}{\partial \Lambda_{ik'}} \Big|_{\Lambda_{ik'}=0}. \tag{4.29}$$

Since $\frac{\partial W_j}{\partial \Lambda_{ij}} = -\frac{\partial W_j}{\partial \mu_j}$ and the externality equals the capacity cost by (4.27), thus

$$v_{ik} - c_{ik} W_k(\mu_k^*, \lambda_k^*) - b_k \geq v_{ik'} - c_{ik'} W_{k'}(\mu_{k'}^*, \lambda_{k'}^*) - b_{k'}. \tag{4.30}$$

As a result, the difference in the individual utility must exceed that in capacity cost, i.e.,

$$v_{ik} - c_{ik} W_k(\mu_k^*, \lambda_k^*) - (v_{ik'} - c_{ik'} W_{k'}(\mu_{k'}^*, \lambda_{k'}^*)) \geq b_k - b_{k'} = p_k - p_{k'}. \tag{4.31}$$

Therefore, by (4.31), if making the price difference equals cost difference, type i customers will actually prefer period k, i.e.,

$$v_{ik} - c_{ik} W_k(\mu_k^*, \lambda_k^*) - p_k \geq v_{ik'} - c_{ik'} W_{k'}(\mu_{k'}^*, \lambda_{k'}^*) - p_{k'}.$$

As we mentioned before, the fixed resource problem can be considered as a special case in which the unit capacity cost is time-invariant. In that particular case, no intertemporal variation exists in unit capacity cost and thus no pricing is required to achieve the incentive compatibility.

Proposition 4.9 shows that only the difference in price matters when coordinating customer individual choices with system efficiency. This relatively weak requirement endows the provider with some confined flexibility in setting the exact prices for TOS. For example, any constant markup on top of the unit capacity cost always serves as a valid incentive compatible scheme. Nonetheless, it is readily shown that under the "price-difference-equal-cost-difference" principle, this constant markup scheme is the only one that can retain the incentive compatibility.

Although the CAFR and CA problems are modeled for different situations about capacity availability, they share many commonalities in their solution structures. First, the optimal customer routing principle is the same. In both models, serving customers of the same type in the same time period appears to be the most efficient way to utilize capacity. Second, the optimal capacity allocation manages congestion subject to capacity cost. In the CA model, the provider freezes his investment in each period's capacity when the marginal capacity cost equals the marginal delay cost. Since the capacity costs are time-varying, marginal delay costs also differ at optimality. In contrast, each unit capacity is equally costly to the provider in the CAFR model. Marginal delay costs thus have no difference across time periods under the optimal routing. Third, differences in capacity costs are essential in achieving incentive compatibility. In particular, unit capacity is identically valuable in the CA model. Hence, no pricing scheme is required.

4.5.4.2 Balking Case

We now relax the no-balking assumption and also assume that the provider controls the arrival rates of different segments. Specifically, the provider solves the problem

$$(\mathbf{CA^b}) \quad \max_{\boldsymbol{\mu}} \; NS(\boldsymbol{\mu}, \boldsymbol{\Lambda}) \;\text{ s.t. }\; (4.20),\; (4.21),\; (4.22),\; (4.23),\; \text{and } (4.24),$$

where (4.20)-(4.24) are the same constraints as in the CAFR$^{\text{b}}$ problem. Constraints (4.20) and (4.21) allow the provider to determine the admission rates of base and strategic customers. Constraints (4.20) and (4.23) ensure all customers receive nonzero expected net utility upon service completion. We once again explore the FB problem to circumvent the technical difficulty caused by the equilibrium constraints (4.23) and (4.24) . We thus ignore these two constraints and the nonzero net utility constraints, which gives rise to the following FB problem

$$(\mathbf{CA^b_{fb}}) \quad \max_{\boldsymbol{\mu},\boldsymbol{\Lambda},\boldsymbol{\Lambda}^b} \; NS(\boldsymbol{\mu}, \boldsymbol{\Lambda}) \text{ s.t. } \; \sum_{j=1}^{n} \Lambda_{ij} \le \overline{\Lambda}_i,\; \lambda_j = \Lambda_j^b + \sum_{i=1}^{m} \Lambda_{ij} < \mu_j,\; \Lambda_{ij} \ge 0 \text{ and } \; 0 \le \Lambda_j^b \le \overline{\Lambda}_j^b.$$

Although the feasible region of the $\text{CA}^{\text{b}}_{\text{fb}}$ problem is not compact, we can apply Proposition 4.7 and reduce the decision variables to $\mathbf{\Lambda}$ and $\mathbf{\Lambda}^b$ only. Thereby, we have an alternative representation of the $\text{CA}^{\text{b}}_{\text{fb}}$ problem

$$\max_{\mathbf{\Lambda},\mathbf{\Lambda}^b} \; NS(\boldsymbol{\mu}^*, \mathbf{\Lambda}) \text{ s.t. } \sum_{j=1}^{n} \Lambda_{ij} \le \overline{\Lambda}_i, \; \Lambda_{ij} \ge 0 \text{ and } \; 0 \le \Lambda^b_j \le \overline{\Lambda}^b_j,$$

where $NS(\boldsymbol{\mu}^*, \mathbf{\Lambda})$ is defined in (4.28). Notice that Propositions 4.8 holds for any given $\mathbf{\Lambda}$, in particular, when $\mathbf{\Lambda} = \mathbf{\Lambda}^*$. In consequence, the incentive compatibility in Proposition 4.9 can also be accomplished when $\mathbf{\Lambda} = \mathbf{\Lambda}^*$ and $\mathbf{\Lambda}^b = \mathbf{\Lambda}^{b*}$. Then, the only issue we need to address is how to induce the optimal arrival rates.

Proposition 4.10 (Pricing for Optimal Arrivals) *Charging for time period j the capacity cost b_j induces the optimal customer arrival rates and TOS choices.*

Recall that any constant markup on the capacity cost would establish the incentive compatibility. However, zero markup leads to the optimal arrival rates to the system.

Let us concentrate on strategic customers to articulate the result. Similar analysis can be applied to base customers as well. Consider the marginal value of serving type i customers. By Proposition (4.8), all type i customers are served in one period, say period j^*. Under the optimal capacity allocation (4.26), the marginal value of serving type i customers is

$$\frac{\partial NS^*}{\partial \Lambda_i} = \frac{\partial NS^*}{\partial \Lambda_{ij^*}} = v_{ij^*} - c_{ij^*} W^*_{j^*} - C_{j^*}(\Lambda_i) \frac{\partial W^*_{j^*}}{\partial \Lambda_{ij^*}} - b_{j^*} \frac{\partial \mu^*_{j^*}}{\partial \Lambda_{ij^*}}$$

where $\mu^*_{j^*} = \mu^*_{j^*}(\Lambda_i)$, $W^*_{j^*}(\Lambda_i) = W_{j^*}(\mu^*_{j^*} \mid \lambda_{j^*})$ and $C_{j^*}(\Lambda_i) = C_{j^*}(\lambda_{j^*}(\Lambda_i))$ are all functions of Λ_i and NS^* is a shorthand of $NS(\boldsymbol{\mu}^*, \mathbf{\Lambda})$ in (4.28). The first two terms $v_{ij^*} - c_{ij^*} W^*_{j^*}$ represent the individual net utility. The third one captures the externality and the last one exhibits the additional capacity cost to accommodate additional customers. Serving more customers increases externality inflicted on others. Yet, the corresponding optimal capacity will also be increased. The net effect from the last two terms turns out to be only dependent on the unit capacity cost, i.e.,

$$\frac{\partial NS^*}{\partial \Lambda_i} = v_{ij^*} - c_{ij^*} W^*_{j^*} - b_{j^*}.$$

At the optimal solution, it must be that $\partial NS^*/\partial \Lambda_i \ge 0$ for $i = 1, \ldots, m$, which implies net utilities that type i customers receive, $v_{ij^*} - c_{ij^*} W^*_{j^*}$, are non-negative. Moreover, it can be shown that $\partial NS^*/\partial \Lambda_i = 0$ only when $\Lambda^*_i < \overline{\Lambda}_i$. Thus, a price b_j may effectively control

the admission rate. Since the prices for TOS exactly match the associated capacity costs, incentive compatibility retains. We note that, like Proposition 4.5, Proposition 4.10 hinges on the assumption that the service time follows an exponential distribution. However, as Dewan and Mendelson (1990) observe, setting prices at marginal costs could be a very good "rule of thumb" for other non-exponential distributions, e.g., Erlang-k distribution.

We can also relate Proposition 4.10 to the CAFR model. Because each unit capacity is equally costly to the provider in that model, a uniform price can induce both optimal arrivals and incentive compatibility.

4.6 Conclusion and Future Research Directions

This paper puts endogeneity between capacity decisions and arrival patterns in perspective. We first model customer TOS choices for a facility that is available over a horizon of multiple time periods. In addition to heterogeneous delay sensitivities, we also allow for heterogeneous preferences on TOS and argue that differences in intertemporal delays may affect customer TOS choices. Such differences further result in the endogeneity between capacity decisions and the intertemporal arrival patten, which is ignored in previous literature. We then consider the optimal capacity allocation problem under customer self-interested TOS choices when taking into account this endogeneity. We find that for a provider with a fixed total capacity without balking, the socially optimal capacity allocation is also incentive compatible. However, if capacity costs are varying over time, the optimal capacity allocation has to simultaneously manage both delay cost and pecuniary cost. Therefore, it is less effective in controlling impacts of congestion. We show that charging prices that equal the unit capacity costs can induce both the incentive compatibility and optimal arrival rates.

4.6.1 Ongoing Work and Potential Future Work

Our analysis and results point to further research directions. We outline three of them here.

First, although time-dependent pricing has its advantage in retaining incentive compatibility, it might be difficult to implement in many scenarios. This gives rise to the necessity to consider the capacity allocation problem in the absence of pricing. In this case, the solution to the FB problem need not be incentive compatible. This raises two questions, (i) under what conditions is incentive-compatibility satisfied, and (ii) in other

cases, how should the provider set capacity levels, compared to the FB solution.

Second, we use system welfare as the performance measure. This is appropriate for public services, but does not well represent private service providers, who aim to maximize their revenues. It is of interest to see how the change in objective from social optimality to provider's profit affects the results.

Third, our results demonstrate the value of studying individual-level service records in order to estimate customer characteristics, e.g., their TOS preferences and delay sensitivities. This information might be practically useful for a service facility, e.g., a call center, who attempts to adjust staffing in the long run. Specifically, because historical arrival data only reflect the demand response to specific past staffing levels. Therefore, historic demand data alone may be misleading for the purpose of predicting future arrivals after capacity adjustments. This type of variability is in fact predictable from historical data with a more sophisticated forecasting framework built on our model.

4.7 Appendix

4.7.1 Proofs

Proof of Proposition 4.1. We show the existence of an equilibrium by constructing an arrival rate matrix $\mathbf{\Lambda}$ that satisfy (EC1) and (EC2) for every type of strategic customers.

Let us first fix the capacity vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$. Following the technique in Beckmann, McGuire and Winstern (1956), we consider the ancillary maximization problem

$$\max_{\mathbf{\Lambda}} \quad \phi(\mathbf{\Lambda}) := \sum_{j=1}^{n} \sum_{i=1}^{m} \int_0^{\sum_{t=1}^{m} \Lambda_{tj}} (v_{ij} - c_{ij} W_j(x))\, dx \tag{4.32}$$

$$\text{s.t.} \quad \sum_{j=1}^{n} \Lambda_{ij} = \Lambda_i, i = 1, \ldots, m, \text{ and } \Lambda_{ij} \geq 0,$$

in decision variables $\mathbf{\Lambda} = (\Lambda_{ij})$, $i = 1, \ldots, m$ and $j = 1, \ldots, n$, where $W_j(x) = \left(\mu_j - \Lambda_j^b - x\right)^{-1}$. The feasible region of problem (4.32) is apparently convex. We will further show that the objective function is concave in $\mathbf{vec}(\mathbf{\Lambda}^T) = (\Lambda_{11}, \ldots, \Lambda_{m1}, \ldots, \Lambda_{1j}, \ldots, \Lambda_{mj}, \ldots, \Lambda_{1n}, \ldots, \Lambda_{mn})^T$, where "$T$" denotes the transpose of a matrix and $\mathbf{vec}(\mathbf{\Lambda}^T)$ is a vectorization operation by stacking $\mathbf{\Lambda}^T$'s columns one by one. Since

$$\frac{\partial \phi}{\partial \Lambda_{ij}} = v_{ij} - c_{ij} W_j \left(\sum\nolimits_{t=1}^{m} \Lambda_{tj}\right) = v_{ij} - c_{ij} \left(\mu_j - \Lambda_j^b - \sum\nolimits_{t=1}^{m} \Lambda_{tj}\right)^{-1},$$

we have for any $i, i' = 1, \ldots, m$ and $j, j' = 1, \ldots, n$,

$$\frac{\partial^2 \phi}{\partial \Lambda_{ij} \partial \Lambda_{i'j'}} = \begin{cases} -c_{ij} \left(-\Lambda_j^b - \sum_{t=1}^m \Lambda_{tj} \right)^{-2} & \text{if } j = j', \\ 0 & \text{if } j \neq j'. \end{cases}$$

Therefore, we obtain the $mn \times mn$ Hessian matrix of ϕ in $\mathbf{vec}(\mathbf{\Lambda}^T)$

$$H(\phi) = \begin{bmatrix} \frac{\partial^2 \phi}{\partial \Lambda_{11}^2} & \cdots & \frac{\partial^2 \phi}{\partial \Lambda_{11} \partial \Lambda_{m1}} & \cdots\cdots & \frac{\partial^2 \phi}{\partial \Lambda_{11} \partial \Lambda_{1n}} & \cdots & \frac{\partial^2 \phi}{\partial \Lambda_{11} \partial \Lambda_{mn}} \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \frac{\partial^2 \phi}{\partial \Lambda_{m1} \partial \Lambda_{11}} & \cdots & \frac{\partial^2 \phi}{\partial \Lambda_{m1}^2} & \cdots\cdots & \frac{\partial^2 \phi}{\partial \Lambda_{m1} \partial \Lambda_{1n}} & \cdots & \frac{\partial^2 \phi}{\partial \Lambda_{m1} \partial \Lambda_{mn}} \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \frac{\partial^2 \phi}{\partial \Lambda_{1n} \partial \Lambda_{11}} & \cdots & \frac{\partial^2 \phi}{\partial \Lambda_{1n} \partial \Lambda_{m1}} & \cdots\cdots & \frac{\partial^2 \phi}{\partial \Lambda_{1n}^2} & \cdots & \frac{\partial^2 \phi}{\partial \Lambda_{1n} \partial \Lambda_{mn}} \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \frac{\partial^2 \phi}{\partial \Lambda_{mn} \partial \Lambda_{11}} & \cdots & \frac{\partial^2 \phi}{\partial \Lambda_{mn} \partial \Lambda_{m1}} & \cdots\cdots & \frac{\partial^2 \phi}{\partial \Lambda_{mn} \partial \Lambda_{1n}} & \cdots & \frac{\partial^2 \phi}{\partial \Lambda_{mn}^2} \end{bmatrix} = \begin{bmatrix} \mathbf{B}_1 & 0 & \cdots & 0 \\ 0 & \mathbf{B}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{B}_n \end{bmatrix},$$

where each diagonal $m \times m$ block

$$\mathbf{B}_j = \frac{-\boldsymbol{c}_j \mathbf{1}^T}{\left(\mu_j - \Lambda_j^b - \sum_{i=1}^m \Lambda_{ij} \right)^2}, \ \boldsymbol{c_j} = (c_{1j}, \ldots, c_{mj})^T, \text{ and } \mathbf{1} = (1, \ldots, 1)^T.$$

Since $\boldsymbol{c}_j \mathbf{1}^T$ is a rank 1 matrix and its only non-zero eigenvalue equals $\boldsymbol{c}_j^T \mathbf{1} > 0$, $\mathbf{B}_j$ is negative semidefinite. Thus, the objective function ϕ is concave and the maximization problem (4.32) is convex. As a result, at least one global maximizer is guaranteed to exist.

We next verify that any maximizer of (4.32) satisfies the TOS equilibrium conditions (EC1) and (EC2) in Definition 1. By the Kuhn-Tucker conditions, any optimal solution to (4.32), which is denoted as $\mathbf{\Lambda}^*$, must satisfy

$$v_{ij} - c_{ij} W_j \left(\sum_{t=1}^m \Lambda_{tj}^* \right) - \omega_i + \gamma_{ij} = 0 \tag{4.33}$$

$$\gamma_{ij} \cdot \Lambda_{ij} = 0 \tag{4.34}$$

for any type i, $i = 1, \ldots, m$, where ω_i and $\gamma_{ij} \geq 0$ are the Lagrange multipliers on constraints $\sum_{j=1}^n \Lambda_{ij} = \Lambda_i$ and $\Lambda_{ij} \geq 0$, respectively.

For each type i, let $\mathcal{P}_i = \{j \mid \Lambda_{ij} > 0\}$ and $\mathcal{Z}_i = \{j \mid \Lambda_{ij} = 0\}$. For any $k, k' \in \mathcal{P}_i$,

$\gamma_{ik} = \gamma_{ik'} = 0$ due to the complementarity condition (4.34) . Therefore, by (4.33),

$$v_{ik} - c_{ik} W_k \left(\sum_{t=1}^{m} \Lambda_{tk}^* \right) = v_{ik'} - c_{ik'} W_{k'} \left(\sum_{t=1}^{m} \Lambda_{tk'}^* \right) = \omega_i. \tag{4.35}$$

For $k \in \mathcal{P}_i$ and $k' \in \mathcal{Z}_i$, $\gamma_{ik} = 0$ and $\gamma_{ik'} \geq 0$. Hence, from (4.33),

$$v_{ik} - c_{ik} W_k \left(\sum_{t=1}^{m} \Lambda_{tk}^* \right) = \omega_i \geq \omega_i - \gamma_{ik'} = v_{ik'} - c_{ik'} W_{k'} \left(\sum_{t=1}^{m} \Lambda_{tk'}^* \right). \tag{4.36}$$

Equations(4.35) and (4.36) coincide with (EC1) and (EC2) in Definition 1, which indicates any maximizer of (4.32) is an equilibrium arrival rate matrix. ■

Proof of Proposition 4.2. Let $\mathbf{\Lambda}^*$ and $\widehat{\mathbf{\Lambda}}^*$ be two distinct equilibria. Thus, they are both maximizers of (4.3). Consider all convex combinations of $\mathbf{\Lambda}^*$ and $\widehat{\mathbf{\Lambda}}^*$ in the form of $t\mathbf{\Lambda}^* + (1-t)\widehat{\mathbf{\Lambda}}^*$ for $t \in [0,1]$. Obviously, $t\mathbf{\Lambda}^* + (1-t)\widehat{\mathbf{\Lambda}}^*$ is a feasible solution. Recall that we have showed the concavity of ϕ in the proof of Proposition 4.1. Therefore, we have

$$\phi(t\mathbf{\Lambda}^* + (1-t)\widehat{\mathbf{\Lambda}}^*) \geq t\phi(\mathbf{\Lambda}^*) + (1-t)\phi(\widehat{\mathbf{\Lambda}}^*).$$

The inequality cannot be strict, since $\mathbf{\Lambda}^*$ and $\widehat{\mathbf{\Lambda}}^*$ are maximizers of (4.3). Therefore, it must be

$$\phi(t\mathbf{\Lambda}^* + (1-t)\widehat{\mathbf{\Lambda}}^*) = t\phi(\mathbf{\Lambda}^*) + (1-t)\phi(\widehat{\mathbf{\Lambda}}^*), \tag{4.37}$$

for any $t \in [0,1]$. Recall that we have shown $\mathbf{B}_j$ is negative semidefinite in the proof of Proposition 4.1, then each summand of ϕ is concave. Therefore, equation (4.37) implies that every summand $\displaystyle\int_0^y (v_{ij} - c_{ij} W_j(x))\, dx$ must be linear between $\sum_{t=1}^m \Lambda_{tj}^*$ and $\sum_{t=1}^m \widehat{\Lambda}_{tj}^*$. Otherwise, (4.37) cannot hold for all $t \in [0,1]$. This further indicates $v_{ij} - c_{ij} W_j(x)$ is a constant between $\sum_{t=1}^m \Lambda_{tj}^*$ and $\sum_{t=1}^m \widehat{\Lambda}_{tj}^*$. At last, since the capacity μ_j is fixed, we obtain $\lambda_j^* = \widehat{\lambda}_j^*$ and $\sum_{i=1}^m \Lambda_{ij}^* = \sum_{i=1}^m \widehat{\Lambda}_{ij}^*$ as well. ■

Proof of Proposition 4.3. The objective function is strictly convex in $\boldsymbol{\mu}$. Thus, the problem is strictly convex and allows an unique optimal solution. By the Kuhn-Tucker conditions, an optimal capacity vector $\boldsymbol{\mu}^* = (\mu_1^*, \ldots, \mu_n^*)$ must satisfy

$$C_1(\lambda_1) \frac{\partial W_1}{\partial \mu_1} \bigg|_{\mu_1 = \mu_1^*} = C_1(\lambda_2) \frac{\partial W_2}{\partial \mu_2} \bigg|_{\mu_2 = \mu_2^*} = \cdots = C_n(\lambda_n) \frac{\partial W_n}{\partial \mu_n} \bigg|_{\mu_n = \mu_n^*} = \omega\mu,$$

where $\dfrac{\partial W_j}{\partial \mu_j} \bigg|_{\mu_j = \mu_j^*} = -(\mu_j^* - \lambda_j)^{-2}$, $j = 1, \ldots, n$, and ω_μ is the Lagrange multiplier on the

constraint $\sum_{j=1}^{n} \mu_j = \mu$. For $j = 1, \ldots, n-1$,

$$C_j(\lambda_j)\frac{\partial W_j}{\partial \mu_j}\bigg|_{\mu_j=\mu_j^*} = C_n(\lambda_n)\frac{\partial W_n}{\partial \mu_n}\bigg|_{\mu_n=\mu_n^*} \iff \frac{C_j(\lambda_j)}{(\mu_j^*-\lambda_j)^2} = \frac{C_n(\lambda_n)}{\left(\mu - \sum_{k=1}^{n-1}\mu_k^* - \lambda_n\right)^2}$$

$$\iff \mu_j^* = \lambda_j + \frac{\sqrt{C_j(\lambda_j)}}{\sqrt{C_n(\lambda_n)}}\left(\mu - \sum_{k=1}^{n-1}\mu_k^* - \lambda_n\right). \quad (4.38)$$

Taking a summation of (4.38) over j, we have

$$\sum_{j=1}^{n-1}\mu_j^* = \sum_{j=1}^{n-1}\lambda_j + \frac{\sum_{j=1}^{n-1}\sqrt{C_j(\lambda_j)}}{\sqrt{C_n(\lambda_n)}}\left(\mu - \sum_{k=1}^{n-1}\mu_k^* - \lambda_n\right). \quad (4.39)$$

Note that $\sum_{j=1}^{n-1}\mu_j^*$ and $\sum_{k=1}^{n-1}\mu_k^*$ are the same variable. Thus, from (4.39),

$$\sum_{k=1}^{n-1}\mu_k^* = \frac{\sqrt{C_n(\lambda_n)}\sum_{j=1}^{n-1}\lambda_j + \sum_{j=1}^{n-1}\sqrt{C_j(\lambda_j)}\,(\mu-\lambda_n)}{\sum_{j=1}^{n}\sqrt{C_j(\lambda_j)}}. \quad (4.40)$$

At last, substitute (4.40) back to (4.38). Then, we obtain

$$\mu_j^* = \lambda_j + \frac{\sqrt{C_j(\lambda_j)}}{\sum_{k=1}^{n}\sqrt{C_k(\lambda_k)}}\left(\mu - \sum_{j=k}^{n}\lambda_k\right),$$

after algebraic simplification. ■

Proof of Proposition 4.4. We will show the result by contradiction. Let $\mathbf{\Lambda}^*$ be the optimal customer routing decision and suppose that type t customers are served in more than one periods, i.e., there are at least two entries of the row vector $\mathbf{\Lambda}_t^* = (\Lambda_{t1}^*, \Lambda_{t2}^*, \ldots, \Lambda_{tn}^*)$ strictly positive. Without loss of generality, let us further assume $\Lambda_{t1}, \Lambda_{t2} > 0$. We will show that $S(\boldsymbol{\mu}^*, \mathbf{\Lambda}^*)$ could be improved by routing all type t customers to either period 1 or 2.

Denote $\Lambda_{t1}^* + \Lambda_{t2}^* = \widehat{\Lambda}_t$ and rewrite (4.52) as a function of Λ_{t1}^*

$$S(\boldsymbol{\mu}^*, \mathbf{\Lambda}^*) = \sum_{j=1}^{n} v_j^b \Lambda_j^b + \sum_{j=1}^{n}\sum_{i\neq t} v_{ij}\Lambda_{ij}^* + \sum_{j=3}^{n} v_{tj}\Lambda_{tj}^* + v_{t1}\Lambda_{t1}^* + v_{t2}\left(\widehat{\Lambda}_t - \Lambda_{t1}^*\right) \quad (4.41)$$

$$- \frac{\left(\sqrt{c_1^b\Lambda_1^b + c_{t1}\Lambda_{t1}^* + \sum_{i\neq t} c_{i1}\Lambda_{i1}^*} + \sqrt{c_2^b\Lambda_2^b + c_{t2}\left(\widehat{\Lambda}_t - \Lambda_{t1}^*\right) + \sum_{i\neq t} c_{i2}\Lambda_{i2}^*} + \sum_{j=3}^{n}\sqrt{c_j^b\Lambda_j^b + \sum_{i=1}^{m} c_{ij}\Lambda_{ij}^*}\right)^2}{\mu - \sum_{j=1}^{n}\Lambda_j^b - \sum_{i=1}^{m}\Lambda_i}.$$

We aim to establish the convexity of $S(\boldsymbol{\mu}^*, \boldsymbol{\Lambda}^*)$ in Λ^*_{t1}. Then, $S(\boldsymbol{\mu}^*, \boldsymbol{\Lambda}^*)$ can be improved by setting either $\Lambda^*_{t1} = 0$ or $\widehat{\Lambda}_t$. For convenience, define

$$K_1 = c^b_1\Lambda^b_1 + \sum_{i\neq t} c_{i1}\Lambda^*_{i1},\ K_2 = c^b_2\Lambda^b_2 + c_{t2}\widehat{\Lambda}_t + \sum_{i\neq t} c_{i2}\Lambda^*_{i2},\ \text{and}\ K_3 = \sum_{j=3}^{n}\sqrt{c^b_j\Lambda^b_j + \sum_{i=1}^{m} c_{ij}\Lambda^*_{ij}}.$$

From (4.41), we see that $S(\boldsymbol{\mu}^*, \boldsymbol{\Lambda})$ is convex in Λ_{k1} if the numerator of the last term in (4.41)

$$\begin{aligned} f(\Lambda_{k1}) &:= \left(\sqrt{c_{t1}\Lambda^*_{t1} + K_1} + \sqrt{-c_{t1}\Lambda^*_{t1} + K_2} + K_3\right)^2 \\ &= K_1 + K_2 + K_3^2 + 2K_3\left(\sqrt{c_{t1}\Lambda^*_{t1} + K_1} + \sqrt{-c_{t1}\Lambda^*_{t1} + K_2}\right) \\ &\quad +2\sqrt{c_{t1}\Lambda^*_{t1} + K_1}\sqrt{-c_{t1}\Lambda^*_{t1} + K_2} \end{aligned}$$

is concave in Λ^*_{t1}. Since $c_{t1}\Lambda_{t1} + K_1 > 0$ and $-c_{t1}\Lambda_{t1} + K_2 > 0$, it is easy to verify $\sqrt{c_{t1}\Lambda^*_{t1} + K_1}$, $\sqrt{-c_{t1}\Lambda^*_{t1} + K_2}$, and the product $\sqrt{c_{t1}\Lambda^*_{t1} + K_1}\sqrt{-c_{t1}\Lambda^*_{t1} + K_2}$ are all concave in Λ^*_{t1}.Therefore, $S(\boldsymbol{\mu}^*, \boldsymbol{\Lambda}^*)$ is convex in Λ^*_{t1}, which indicates ceteris paribus, $S(\boldsymbol{\mu}^*, \boldsymbol{\Lambda}^*)$ can be improved by choosing $\Lambda^*_{t1} = 0$ or $\widehat{\Lambda}_t$, i.e., pooling type t strategic customers who are served in periods 1 or 2 to one of the two periods can improve the social welfare $S(\boldsymbol{\mu}^*, \boldsymbol{\Lambda}^*)$. This contradicts with the fact that $\boldsymbol{\Lambda}^*$ is the optimal customer routing decision. Therefore, strategic customers of the same type must be routed to exactly one period at optimality. ■

Proof of Proposition 4.5. Let us consider the Lagrangian for the FB problem

$$\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{\omega}, \omega_\mu, \boldsymbol{\nu}, \boldsymbol{\Gamma}) = S(\boldsymbol{\mu}, \boldsymbol{\Lambda}) + \sum_i w_i\left(\Lambda_i - \sum_{j=1}^{n}\Lambda_{ij}\right) + \omega_\mu\left(\mu - \sum_j \mu_j\right) + \sum_j \nu_j(\mu_j - \lambda_j) + \sum_{i,j}\gamma_{ij}\Lambda_{ij},$$

where $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_m)$, ω_μ, $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_n) \geq 0$, and $\boldsymbol{\Gamma} = (\gamma_{ij}) \geq 0$ are Lagrange multipliers for the corresponding constraints. By the Kuhn-Tucker conditions, the optimal solution to the FB problem must satisfy

$$\frac{\partial \mathcal{L}}{\partial \Lambda_{ij}} = v_{ij} - c_{ij}W_j - \left(c^b_j\Lambda^b_j + \sum_{t=1}^{m} c_{tj}\Lambda_{tj}\right)\frac{\partial W_j}{\partial \Lambda_{ij}} - \omega_i - \nu_j + \gamma_{ij} = 0,\ \forall\ j = 1, \ldots, n \text{ and } i = 1, \ldots, \quad (4.42)$$

$$\frac{\partial \mathcal{L}}{\partial \mu_j} = -\left(c^b_j\Lambda^b_j + \sum_{t=1}^{m} c_{tj}\Lambda_{tj}\right)\frac{\partial W_j}{\partial \mu_j} - \omega_\mu + \nu_j = 0,\ \forall\ j = 1, \ldots, n \quad (4.43)$$

$$\nu_j \cdot (\mu_j - \lambda_j) = 0,\ \forall\ j = 1, \ldots, n \quad (4.44)$$

$$\gamma_{ij} \cdot \Lambda_{ij} = 0,\ \forall\ j = 1, \ldots, n \text{ and } i = 1, \ldots, \quad (4.45)$$

Since $\dfrac{\partial W_j}{\partial \Lambda_{ij}} = -\dfrac{\partial W_j}{\partial \mu_j}$ and $\dfrac{\partial \mathcal{L}}{\partial \mu_j} = 0$ by (4.43), we have

$$\frac{\partial \mathcal{L}}{\partial \Lambda_{ij}} = v_{ij} - c_{ij}W_j + \left(c_j^b \Lambda_j^b + \sum_{t=1}^m c_{tj}\Lambda_{tj}\right)\frac{\partial W_j}{\partial \mu_j} - \omega_i - \nu_j + \gamma_{ij} = v_{ij} - c_{ij}W_j - \omega_\mu - \omega_i + \gamma_{ij} = 0.$$

From Proposition 4.4, for each type i, only one element of $\mathbf{\Lambda}_i = (\Lambda_{i1}, \ldots, \Lambda_{in})$ is strictly positive and equal to Λ_i, say, $\Lambda_{ik} = \Lambda_i$. Then, $\Lambda_{ik'} = 0$ for $k' \neq k$. Moreover, the complementarity condition (4.45) implies

$$\gamma_{ik} = 0 \text{ and } \gamma_{ik'} \geq 0 \text{ for } k' \neq k.$$

As a result, for type i customers, $i = 1, \ldots, m$,

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \Lambda_{ik}} = \frac{\partial \mathcal{L}}{\partial \Lambda_{ik'}} = 0 \quad &\Longleftrightarrow \quad v_{ik} - c_{ik}W_k - \omega_\mu - \omega_i = v_{ik'} - c_{ik'}W_{k'} - \omega_\mu - \omega_i + \gamma_{ik'} \\ &\Longleftrightarrow \quad v_{ik} - c_{ik}W_k \geq v_{ik'} - c_{ik'}W_{k'}, \end{aligned}$$

which demonstrates that it is actually optimal for type i customers to choose period k than any other period and they have no incentive to deviate from the provider's FB solution. ■

Proof of Proposition 4.6. We first consider the marginal value of serving an additional amount of type i customers. Assume the total capacity is optimally allocated according to (4.12) for known arrival rates $\mathbf{\Lambda}$ and $\mathbf{\Lambda}^b$. Moreover, by Proposition 4.4, all type i customers must be served in one time period, say period j^*. Then, expected delay in every period W_j^*, $j = 1, \ldots, n$, and period j^*'s unit delay cost C_{j^*} are all functions of Λ_i. Hence, we rewrite $S(\boldsymbol{\mu}^*, \mathbf{\Lambda})$ in (4.14) as

$$S(\boldsymbol{\mu}^*, \mathbf{\Lambda}) = \sum_{j=1}^n v_j^b \Lambda_j^b + \sum_{j=1}^n \sum_{i=1}^m v_{ij}\Lambda_{ij} - C_{j^*}(\Lambda_i)W_{j^*}^*(\Lambda_i) - \sum_{j \neq j^*} C_j W_j^*(\Lambda_i),$$

where $W_j^*(\Lambda_j) = W_j(\mu_j^*(\Lambda_i) \mid \lambda_j)$, $j = 1, \ldots, n$, and $C_{j^*}(\Lambda_i) = C_{j^*}(\lambda_{j^*}(\Lambda_i))$. Since type i customers are presumably being served in period j^*,

$$\begin{aligned} \frac{\partial S^*}{\partial \Lambda_i} = \frac{\partial S^*}{\partial \Lambda_{ij^*}} &= v_{ij^*} - \frac{\partial C_{j^*}}{\partial \Lambda_i}W_{j^*}^*(\Lambda_i) - C_{j^*}(\Lambda_i)\frac{\partial W_{j^*}^*}{\partial \Lambda_i} - \sum_{j \neq j^*} C_j \frac{\partial W_j^*}{\partial \Lambda_i} \\ &= v_{ij^*} - c_{ij}W_{j^*}^* - C_{j^*}(\Lambda_i)\frac{\partial W_{j^*}^*}{\partial \Lambda_i} - \sum_{j \neq j^*} C_j \frac{\partial W_j^*}{\partial \Lambda_i}, \end{aligned} \tag{4.46}$$

where S^* is a shorthand for $S(\boldsymbol{\mu}^*, \boldsymbol{\Lambda})$. Recall from (4.12),

$$\begin{aligned}
\frac{\partial W_{j^*}^*}{\partial \Lambda_i} &= \frac{\partial}{\partial \Lambda_i}\left(\frac{1}{\mu_{j^*}^* - \lambda_{j^*}}\right) = \frac{\partial}{\partial \Lambda_i}\left(\frac{\sum_{k=1}^n \sqrt{C_k}}{\sqrt{C_{j^*}(\Lambda_i)}} \cdot \frac{1}{\mu - \sum_{k=1}^n \lambda_k}\right) \\
&= \frac{\partial}{\partial \Lambda_i}\left(\frac{\sum_{k=1}^n \sqrt{C_k}}{\sqrt{C_{j^*}(\Lambda_i)}}\right)\frac{1}{\mu - \sum_{k=1}^n \lambda_k} + \frac{\sum_{k=1}^n \sqrt{C_k}}{\sqrt{C_{j^*}(\Lambda_i)}}\frac{\partial}{\partial \Lambda_i}\left(\frac{1}{\mu - \sum_{k=1}^n \lambda_k}\right) \\
&= \frac{\partial}{\partial \Lambda_i}\left(\frac{1}{\sqrt{C_{j^*}(\Lambda_i)}}\right)\frac{\sum_{k\neq j^*}^n \sqrt{C_k}}{\mu - \sum_{k=1}^n \lambda_k} + \frac{\sum_{k=1}^n \sqrt{C_k}}{\sqrt{C_{j^*}(\Lambda_i)}}\frac{\partial}{\partial \Lambda_i}\left(\frac{1}{\mu - \sum_{k=1}^n \lambda_k}\right) \qquad (4.47)
\end{aligned}$$

and

$$\begin{aligned}
\frac{\partial W_{j}^*}{\partial \Lambda_i} &= \frac{\partial}{\partial \Lambda_i}\left(\frac{1}{\mu_{j^*}^* - \lambda_{j}}\right) = \frac{\partial}{\partial \Lambda_i}\left(\frac{\sum_{k=1}^n \sqrt{C_k}}{\sqrt{C_j}} \cdot \frac{1}{\mu - \sum_{k=1}^n \lambda_k}\right) \\
&= \frac{\partial}{\partial \Lambda_i}\left(\frac{\sum_{k=1}^n \sqrt{C_k}}{\sqrt{C_j}}\right)\frac{1}{\mu - \sum_{k=1}^n \lambda_k} + \frac{\sum_{k=1}^n \sqrt{C_k}}{\sqrt{C_j}}\frac{\partial}{\partial \Lambda_i}\left(\frac{1}{\mu - \sum_{k=1}^n \lambda_k}\right) \\
&= \frac{1}{\sqrt{C_j}}\frac{\partial}{\partial \Lambda_i}\left(\sqrt{C_{j^*}(\Lambda_i)}\right)\frac{1}{\mu - \sum_{k=1}^n \lambda_k} + \frac{\sum_{k=1}^n \sqrt{C_k}}{\sqrt{C_j}}\frac{\partial}{\partial \Lambda_i}\left(\frac{1}{\mu - \sum_{k=1}^n \lambda_k}\right) \qquad (4.48)
\end{aligned}$$

Substitue (4.47) and (4.48) to (4.46),

$$\begin{aligned}
\frac{\partial S^*}{\partial \Lambda_i} &= v_{ij^*} - c_{ij^*}W_{j^*}^* - \frac{\partial}{\partial \Lambda_i}\left(\frac{1}{\sqrt{C_{j^*}(\Lambda_i)}}\right)\frac{C_{j^*}(\Lambda_i)\sum\limits_{k\neq j^*}^n \sqrt{C_k}}{\mu - \sum_{k=1}^n \lambda_k} - \sqrt{C_{j^*}(\Lambda_i)}\sum_{k=1}^n \sqrt{C_k}\frac{\partial}{\partial \Lambda_i}\left(\frac{1}{\mu - \sum_{k=1}^n \lambda_k}\right) \\
&\quad - \sum_{j\neq j^*}^n \left(\frac{\partial}{\partial \Lambda_i}\left(\sqrt{C_{j^*}(\Lambda_i)}\right)\frac{\sqrt{C_j}}{\mu - \sum_{k=1}^n \lambda_k} + \sqrt{C_j}\sum_{k=1}^n \sqrt{C_k}\frac{\partial}{\partial \Lambda_i}\left(\frac{1}{\mu - \sum_{k=1}^n \lambda_k}\right)\right) \\
&= v_{ij^*} - c_{ij^*}W_{j^*}^* - \left(\sum_{k=1}^n \sqrt{C_k}\right)^2 \frac{\partial}{\partial \Lambda_i}\left(\frac{1}{\mu - \sum_{k=1}^n \lambda_k}\right) \\
&\quad - \frac{C_{j^*}(\Lambda_i)\sum_{k\neq j^*}^n \sqrt{C_k}}{\mu - \sum_{k=1}^n \lambda_k}\frac{\partial}{\partial \Lambda_i}\left(\frac{1}{\sqrt{C_{j^*}(\Lambda_i)}}\right) - \frac{\sum_{k\neq j^*}^n \sqrt{C_k}}{\mu - \sum_{k=1}^n \lambda_k}\frac{\partial}{\partial \Lambda_i}\left(\sqrt{C_{j^*}(\Lambda_i)}\right) \\
&= v_{ij^*} - c_{ij^*}W_{j^*}^* - \left(\sum_{k=1}^n \sqrt{C_k}\right)^2 \frac{\partial}{\partial \Lambda_i}\left(\frac{1}{\mu - \sum_{k=1}^n \lambda_k}\right) \\
&\quad - \frac{\sum_{k\neq j^*}^n \sqrt{C_k}}{\mu - \sum_{k=1}^n \lambda_k}\left(C_{j^*}(\Lambda_i)\frac{\partial}{\partial \Lambda_i}\left(\frac{1}{\sqrt{C_{j^*}(\Lambda_i)}}\right) + \frac{\partial}{\partial \Lambda_i}\left(\sqrt{C_{j^*}(\Lambda_i)}\right)\right) \\
&= v_{ij^*} - c_{ij^*}W_{j^*}^* - \left(\sum_{k=1}^n \sqrt{C_k}\right)^2 \frac{\partial}{\partial \Lambda_i}\left(\frac{1}{\mu - \sum_{k=1}^n \lambda_k}\right)
\end{aligned}$$

$$-\frac{\sum_{k\neq j^*}^n \sqrt{C_k}}{\mu-\sum_{k=1}^n \lambda_k}\left(-\frac{C_{j^*}(\Lambda_i)}{C_{j^*}(\Lambda_i)}\frac{\partial}{\partial \Lambda_i}\left(\sqrt{C_{j^*}(\Lambda_i)}\right)+\frac{\partial}{\partial \Lambda_i}\left(\sqrt{C_{j^*}(\Lambda_i)}\right)\right)$$

$$= v_{ij^*}-c_{ij^*}W^*_{j^*}-\left(\sum_{k=1}^n \sqrt{C_k}\right)^2 \frac{\partial}{\partial \Lambda_i}\left(\frac{1}{\mu-\sum_{k=1}^n \lambda_k}\right).$$

$$= v_{ij^*}-c_{ij^*}W^*_{j^*}-\left(\sum_{k=1}^n \sqrt{C_k}\right)^2 \left(\mu-\sum_{k=1}^n \lambda_k\right)^{-2}. \tag{4.49}$$

Equation (4.49) shows the marginal value of serving type i customers equals individual utility from that additional customer minus the externality inflicted on the system. At the optimal solution, this marginal value must be non-negative, i.e.,

$$\frac{\partial S^*}{\partial \Lambda_i}\Big|_{\Lambda_i=\Lambda_i^*} = v_{ij^*}-c_{ij^*}W_{j^*}(\mu^*_{j^*}\mid \lambda^*_{j^*})-\left(\sum_{k=1}^n \sqrt{C_k(\lambda_k^*)}\right)^2\left(\mu-\sum_{k=1}^n \lambda_k^*\right)^{-2}\geq 0. \tag{4.50}$$

In particular, if $\frac{\partial S^*}{\partial \Lambda_i}\Big|_{\Lambda_i=\Lambda_i^*}>0$, $\Lambda_i^*=\overline{\Lambda}_i$.[5] Therefore, if charging a price

$$p=\left(\sum_{k=1}^n \sqrt{C_k(\lambda_k^*)}\right)^2\left(\mu-\sum_{k=1}^n \lambda_k^*\right)^{-2}$$

at $\mathbf{\Lambda}=\mathbf{\Lambda}^*$ and $\mathbf{\Lambda}^b=\mathbf{\Lambda}^{b*}$, the provider will serve all type i customers if $\frac{\partial S^*}{\partial \Lambda_i}\Big|_{\Lambda_i=\Lambda_i^*}>0$ or only a fraction $\Lambda_i^*\leq\overline{\Lambda}_i$ if $\frac{\partial S^*}{\partial \Lambda_i}\Big|_{\Lambda_i=\Lambda_i^*}=0$. And all served strategic customers receive non-negative utilities after paying the admission fee p according to (4.50).

Similarly, we can derive the marginal change in social welfare by serving base customers of period j,

$$\frac{\partial S^*}{\partial \Lambda_j^b}=v_j^b-c_j^bW_j^*-\left(\sum_{k=1}^n \sqrt{C_k}\right)^2\left(\mu-\sum_{k=1}^n \lambda_k\right)^{-2}. \tag{4.51}$$

The optimal arrival rates can be achieved by the same argument. ■

Proof of Proposition 4.7. Since $W_j(\mu_j\mid\lambda_j)=(\mu_j-\lambda_j)^{-1}$ is strictly convex in $\mu_j>\lambda_j$ for a fixed $\lambda_j\geq 0$, the objective function $\sum_{j=1}^n C_jW_j(\mu_j\mid\lambda_j)+\sum_{j=1}^n b_j\mu_j$ is jointly

[5]This result is not as trivial as it seems to be. Since a time period may simultaneously serve multiple types of strategic customers, it is not clear that the provider would exhaust one type before starting to serve another. Let us say it is type i that is being served in period j. As the number of type i customers increases, the marginal value of serving type i customers decreases and may be dominated by that of another type, since strategic customers have different valuations and delay sensitivities. A rigorous proof of the result can be accomplished by invoking the Kuhn-Tucker conditions of the $\text{CAFR}^{\text{b}}_{\text{fb}}$ problem and considering values of the Lagrange multipliers that corresponds to arrival rates of each type. A detailed proof is also available from the authors upon request.

convex in $\boldsymbol{\mu} > \boldsymbol{\lambda}$. By the first order condition, we have for any $j = 1, \ldots, n$,

$$C_j(\lambda_j)\frac{\partial W_j}{\partial \mu_j} + b_j = -\frac{C_j(\lambda_j)}{(\mu_j^* - \lambda_j)^2} + b_j = 0$$

at optimality. Therefore, $\mu_j^* = \lambda_j + \sqrt{C_j(\lambda_j)/b_j}$. ■

*Proof of Proposition 4.8.*We show, by contradiction, that $NS(\boldsymbol{\mu}^*, \boldsymbol{\Lambda})$ can be maximized when only one entry of row $\boldsymbol{\Lambda}_t = (\Lambda_{t1}, \ldots, \Lambda_{tn})$, $t = 1, \ldots, m$, is nonzero.

Let $\boldsymbol{\Lambda}^*$ be the optimal customer routing decision and Suppose that for type t strategic customers, there are at least two strictly positive entries of row $\boldsymbol{\Lambda}_t^*$. Without loss of generality, let us further assume $\Lambda_{t1}^*, \Lambda_{t2}^* > 0$. We will prove that $NS(\boldsymbol{\mu}^*, \boldsymbol{\Lambda}^*)$ could be improved by routing either all type t flexible customers in period 1 to period 2. Let $\widehat{\Lambda}_t := \Lambda_{t1}^* + \Lambda_{t2}^*$ and rewrite (4.28) as

$$\begin{aligned} NS(\boldsymbol{\mu}^*, \boldsymbol{\Lambda}^*) \;=\; & \sum_{j=1}^n v_j^b \Lambda_j^b - \sum_{j=1}^n b_j \Lambda_j^b + \sum_{j=1}^n \sum_{i \neq t} v_{ij} \Lambda_{ij}^* + v_{t1}\Lambda_{t1}^* + v_{t2}\left(\widehat{\Lambda}_t - \Lambda_{t1}^*\right) + \sum_{j=3}^n v_{tj}\Lambda_{tj}^* \\ & - \sum_{j=1}^n \left(b_j \sum_{i \neq t} \Lambda_{ij}^* \right) - b_1 \Lambda_{t1}^* - b_2\left(\widehat{\Lambda}_t - \Lambda_{t1}^*\right) - \sum_{j=3}^n b_j \Lambda_{tj}^* - 2\sum_{j=3}^n \sqrt{b_j \left(c_j^b \Lambda_j^b + \sum_{i=1}^m c_{ij}\Lambda_{ij}^* \right)} \\ & - \sqrt{b_1 \left(c_1^b \Lambda_1^b + \sum_{i \neq t} c_{i1}\Lambda_{i1}^* + c_{t1}\Lambda_{t1}^* \right)} - \sqrt{b_2 \left(c_2^b \Lambda_2^b + \sum_{i \neq t} c_{i2}\Lambda_{i2}^* + c_{t2}\left(\widehat{\Lambda}_t - \Lambda_{t1}^*\right) \right)}. \end{aligned}$$

Note that all radicands are positive and the last two terms are convex in Λ_{t1}^*. Provided that all other terms are linear in Λ_{t1}^*, $NS(\boldsymbol{\mu}^*, \boldsymbol{\Lambda}^*)$ is thus convex in Λ_{t1}. The maximizer must reside on the boundary, which implies $NS(\boldsymbol{\mu}^*, \boldsymbol{\Lambda}^*)$ could be improved by routing all type t strategic customers to either period 1 or period 2. This contradicts with the fact that $\boldsymbol{\Lambda}^*$ is the optimal customer routing decision. Therefore, strategic customers of the same type must be routed to exactly one period at optimality. ■

Proof of Proposition 4.9. Let us consider the Lagrangian for the FB problem

$$\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Lambda}, \boldsymbol{\nu}, \boldsymbol{\Gamma}) = S(\boldsymbol{\mu}, \boldsymbol{\Lambda}) - \sum_{j=1}^n b_j \mu_j + \sum_i w_i \left(\Lambda_i - \sum_{j=1}^n \Lambda_{ij} \right) + \sum_j \nu_j(\mu_j - \lambda_j) + \sum_{i,j} \gamma_{ij}\Lambda_{ij},$$

where $\boldsymbol{v} = (\nu_1, \ldots, \nu_n) \geq 0$, and $\boldsymbol{\Gamma} = (\gamma_{ij}) \geq 0$ are Lagrange multipliers. By the Kuhn-Tucker conditions, the optimal solution must satisfy

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \Lambda_{ij}} = v_{ij} - c_{ij}W_j - \left(c_j^b \Lambda_j^b + \textstyle\sum_{t=1}^m c_{tj}\Lambda_{tj}\right)\frac{\partial W_j}{\partial \Lambda_{ij}} - \omega_i - \nu_j + \gamma_{ij} &= 0,\ \forall\ j = 1, \ldots, n \text{ and } i = 1, \ldots, m \\ \frac{\partial \mathcal{L}}{\partial \mu_j} = -\left(c_j^b \Lambda_j^b + \textstyle\sum_{t=1}^m c_{tj}\Lambda_{tj}\right)\frac{\partial W_j}{\partial \mu_j} - b_j + \nu_j &= 0,\ \forall\ j = 1, \ldots, n \\ \nu_j \cdot (\mu_j - \lambda_j) &= 0,\ \forall\ j = 1, \ldots, n \end{aligned}$$

$$\gamma_{ij} \cdot \Lambda_{ij} \quad = \quad 0, \ \forall \ i, j = 1, \ldots, n \text{ and } i = 1, \ldots, m.$$

Since $\frac{\partial W_j}{\partial \Lambda_{ij}} = -\frac{\partial W_j}{\partial \mu_j}$ and $\frac{\partial \mathcal{L}}{\partial \mu_j} = 0$, we have

$$\frac{\partial \mathcal{L}}{\partial \Lambda_{ij}} = v_{ij} - c_{ij} W_j + \left(c_j^b \Lambda_j^b + \sum\nolimits_{t=1}^{m} c_{tj} \Lambda_{tj} \right) \frac{\partial W_j}{\partial \mu_j} - \omega_i - \nu_j + \gamma_{ij} = v_{ij} - c_{ij} W_j - k_j - \omega_i + \gamma_{ij} = 0.$$

By Proposition 4.4, for each type i, only one element of $\mathbf{\Lambda}_i = (\Lambda_{i1}, \ldots, \Lambda_{in})$ is positive and must equal to Λ_i. Let us say it is $\Lambda_{ik} = \Lambda_i$. Then, $\Lambda_{ik'} = 0$ for $k' \neq k$. Moreover, the complementarity of Lagrange multipliers implies

$$\gamma_{ik} = 0 \text{ and } \gamma_{ik'} \geq 0 \text{ for } k' \neq k.$$

As a result, for type i customers,

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \Lambda_{ik}} = \frac{\partial \mathcal{L}}{\partial \Lambda_{ik'}} = 0 \quad &\Longleftrightarrow \quad v_{ik} - c_{ik} W_k - b_k - \omega_i = v_{ik'} - c_{ik'} W_{k'} - b_{k'} - \omega_i + \gamma_{ik'} \\ &\Longleftrightarrow \quad v_{ik} - c_{ik} W_k - b_k \geq v_{ik'} - c_{ik'} W_{k'} - b_{k'}, \end{aligned}$$

which verifies type i customers have no incentive to deviate to other period at the FB solution. ■

Proof of Proposition 4.10. For any given arrival rate matrix $\mathbf{\Lambda}$, assume the total capacity is optimally allocated according to (4.26). The net value of system welfare and capacity cost can be expressed as

$$NS(\boldsymbol{\mu}^*, \mathbf{\Lambda}) = S(\boldsymbol{\mu}^*, \mathbf{\Lambda}) - \sum_{j=1}^{n} b_j \mu_j^* \quad = \quad \sum_{j=1}^{n} v_j^b \Lambda_j^b + \sum_{j=1}^{n} \sum_{i=1}^{m} v_{ij} \Lambda_{ij} - \sum_{j=1}^{n} C_j W_j^* - \sum_{j=1}^{n} b_j \mu_j^*,$$

which is only a function $\mathbf{\Lambda}$, since $\boldsymbol{\mu}^*$ is specified as (4.26). By Proposition 4.8, all type i customers are served in one time period, say period j^*. Then, we rewrite $NS(\boldsymbol{\mu}^*, \mathbf{\Lambda})$ as

$$NS^* := NS(\boldsymbol{\mu}^*, \mathbf{\Lambda}) \quad = \quad \sum_{j=1}^{n} v_j^b \Lambda_j^b + \sum_{j=1}^{n} \sum_{i=1}^{m} v_{ij} \Lambda_{ij} - C_{j^*} W_{j^*}^* - \sum_{j \neq j^*} C_j W_j^* - b_{j^*} \mu_{j^*}^* - \sum_{j \neq j^*}^{n} b_j \mu_j^*$$

Note that C_{j^*}, $W_{j^*}^*$ and $\mu_{j^*}^*$ are all functions of Λ_i. Thus, the marginal change in net system welfare by serving type i in period j^* is

$$\begin{aligned} \frac{\partial NS^*}{\partial \Lambda_{ij^*}} \quad &= \quad v_{ij^*} - \frac{\partial C_{j^*}}{\partial \Lambda_{ij^*}} W_{j^*}^* - C_{j^*} \frac{\partial W_{j^*}^*}{\partial \Lambda_{ij^*}} - b_{j^*} \frac{\partial \mu_{j^*}^*}{\partial \Lambda_{ij^*}} \\ &= \quad v_{ij^*} - \frac{\partial C_{j^*}}{\partial \Lambda_{ij^*}} W_{j^*}^* - C_{j^*} \frac{\partial}{\partial \Lambda_{ij^*}} \left(\sqrt{b_{j^*} / C_{j^*}} \right) - b_{j^*} \frac{\partial}{\partial \Lambda_{ij^*}} \left(\lambda_{j^*} + \sqrt{C_{j^*} / b_{j^*}} \right) \end{aligned}$$

$$
\begin{aligned}
&= v_{ij^*} - \frac{\partial C_{j^*}}{\partial \Lambda_{ij^*}} W^*_{j^*} - C_{j^*} \frac{\partial}{\partial \Lambda_{ij^*}} \left(\sqrt{b_{j^*}/C_{j^*}} \right) - b_{j^*} \frac{\partial \lambda_{j^*}}{\partial \Lambda_{ij^*}} + b_{j^*} \frac{C_{j^*}}{b_{j^*}} \cdot \frac{\partial}{\partial \Lambda_{ij^*}} \left(\sqrt{b_{j^*}/C_{j^*}} \right) \\
&= v_{ij^*} - c_{ij^*} W^*_{j^*} - b_{j^*},
\end{aligned}
$$

where the second last equation results from

$$
\frac{\partial}{\partial \Lambda_{ij^*}} \left(\sqrt{C_{j^*}/b_{j^*}} \right) = \frac{\partial}{\partial \Lambda_{ij^*}} \left(1 \Big/ \sqrt{b_{j^*}/C_{j^*}} \right) = -\frac{C_{j^*}}{b_{j^*}} \cdot \frac{\partial}{\partial \Lambda_{ij^*}} \left(\sqrt{b_{j^*}/C_{j^*}} \right).
$$

At the optimal solution, the marginal value of serving type i customers must be non-negative, i.e.,

$$
\frac{\partial NS^*}{\partial \Lambda_{ij^*}} \Bigg|_{\Lambda_{ij^*} = \Lambda^*_i} = v_{ij^*} - c_{ij^*} W^*_{j^*} - b_{j^*} \geq 0. \tag{4.52}
$$

In particular, if $\frac{\partial NS^*}{\partial \Lambda_{ij^*}} \Big|_{\Lambda_{ij^*} = \Lambda^*_i} > 0$, $\Lambda^*_i = \overline{\Lambda}_i$. Therefore, if charging a price $p = b_{j^*}$ at $\mathbf{\Lambda} = \mathbf{\Lambda}^*$ and $\mathbf{\Lambda}^b = \mathbf{\Lambda}^{b*}$, the provider will serve all type i customers if $\frac{\partial NS^*}{\partial \Lambda_{ij^*}} \Big|_{\Lambda_{ij^*} = \Lambda^*_i} > 0$ or only a fraction $\Lambda^*_i \leq \overline{\Lambda}_i$ if $\frac{\partial NS^*}{\partial \Lambda_{ij^*}} \Big|_{\Lambda_{ij^*} = \Lambda^*_i} = 0$. And all served strategic customers receive non-negative utilities after paying the admission fee p according to (4.52). Since $p_j = b_j$, the incentive compatibility attains.

We can apply similar arguments to base customers and obtain the same results. ■

Bibliography

Acemoglu, D., K. Bimpikis, A. Ozdaglar. 2009. Price and capacity competition. *Game & Econ. Behav.* **66**(1) 1–26.

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

Aguir, M. S., O. Z. Akşin, F. Karaesmen, Y. Dallery. 2008. On the interaction between retrials and sizing of call centers. *Eur. J. Oper. Res.* **191**(2) 398–408.

Akşin, O. Z., M. Armony, V. Mehrotra. 2007. The modern call center: A multi-disciplinary perspective on operations management research. *Production Oper. Management* **16**(6) 665–688.

Akşin, Z, B. Ata, S. M. Emadi, C.-L. Su. 2013. Structural estimation of callers' delay sensitivity in call centers. *Management Sci.* **59**(12) 2727–2746.

Allon, G., A. Bassamboo. 2011. The impact of delaying the delay announcements. *Oper. Res.* **59**(5) 1198–1210.

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

Allon, G., A. Federgruen. 2007. Competition in service industries. *Oper. Res.* **55**(1) 37–55.

Anand, K. S., K. Girotra. 2007. The strategic perils of delayed differentiation. *Management Sci.* **53**(5) 697–712.

Anupindi, R., L. Jiang. 2008. Capacity investment under postponement strategies, market competition, and demand uncertainty. *Management Sci.* **54**(11) 1876–1890.

Armony, M., C. Maglaras. 2004a. Contact centers with a call-back option and real-time delay information. *Oper. Res.* **52**(4) 527–545.

Armony, M., C. Maglaras. 2004b. On customer contact centers with a call-back option: customer decisions, routing rules, and system design. *Oper. Res.* **52**(2) 271–292.

Arnott, R., A. de Palma, R. Lindsey. 1993. A structural model of peak-period congestion: A traffic bottleneck with elastic demand. *Amer. Econom. Rev.* **83**(1) pp. 161–179.

Beckmann, M. J., C. B. McGuire, C. B. Winsten. 1956. *Studies in the Economics of Transportation*. Yale University Press.

Bell, C. E., S. Jr. Stidham. 1983. Individual versus social optimization in the allocation of customers to alternative servers. *Management Sci.* **29**(7) 831–839.

Berman, O., D. Krass. 2015. Stochastic location models with congestion. G. Laporte, S. Nickel, F. Saldanha da Gama, eds., *Location Science*. Springer International Publishing, 443–486.

Bernstein, F., A. Federgruen. 2004. Dynamic inventory and pricing models for competing retailers. *Nav. Res. Logist.* **51**(2) 258–274.

Bernstein, F., A. Federgruen. 2005. Decentralized supply chains with competing retailers under demand uncertainty. *Management Sci.* **51**(1) 18–29.

Boccard, N., X. Wauthy. 2000. Bertrand competition and cournot outcomes: Further results. *Econom. Let.* **68**(3) 279–285.

Bradford, R. M. 1996. Pricing, routing, and incentive compatibility in multiserver queues. *Eur. J. Oper. Res.* **89**(2) 226 – 236.

Brown, L., N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, L. Zhao. 2005. Statistical analysis of a telephone call center: A queueing-science perspective. *J. Amer. Statist. Assoc.* **100** 36–50.

Cachon, G. P., R. Swinney. 2009. Purchasing, pricing, and quick response in the presence of strategic consumers. *Management Sci.* **55**(3) 497–511.

Caro, F., V. Martínez-de-Albéniz. 2010. The impact of quick response in inventory-based competition. *Manufacturing Service Oper. Management* **12**(3) 409–429.

Chen, H., M. Frank. 2004. Monopoly pricing when customers queue. *IIE Trans.* **36**(6) 569–581.

Chen, X., D. Simchi-levi. 2012. Pricing and inventory management. O. Ozer, R. Phillips, eds., *The Oxford Handbook of Pricing Management*. Oxford University Press, UK.

Council on Physician and Nurse Supply. 2007. 2007 national physician and nurse supply survey. conducted by AMN Healthcare. `http://www.physiciannursesupply.com/Articles/council-survey-2007.pdf`.

Cui, S., X. Su, S. K. Veeraraghavan. 2014. A model of rational retrials in queues. Available at SSRN: `http://ssrn.com/abstract=2344510`.

Cui, S., S. Veeraraghavan. 2014. Blind queues: The impact of consumer beliefs on revenues and congestion. Working Paper, The Wharton School, University of Pennsylvania, Philadelphia, PA.

de Palma, A., M. Fosgerau. 2011. Dynamic and static congestion models: a review. A. de Palma, R. Lindsey, E. Quinet, R. Vickerman, eds., *A Handb. Transp. Econ.*, chap. 9. Edward Elgar Pub, 188–212.

de Palma, A., R. Lindsey. 2011. Traffic congestion pricing methodologies and technologies. *Transportation Research Part C: Emerging Technologies* **19**(6) 1377 – 1399.

de Véricourt, F., Y.-P. Zhou. 2005. Managing response time in a call-routing problem with service failure. *Oper. Res.* **53**(6) 968–981.

Debo, L., S. Veeraraghavan. 2014. Equilibrium in queues under unknown service times and service value. *Oper. Res.* **62**(1) 38–57.

Debo, L. G., C. Parlour, U. Rajan. 2012. Signaling quality via queues. *Management Sci.* **58**(5) 876–891.

Desmarais, M. n.d. The call center's main purpose is to retain customers. `http://www.sqmgroup.com/retaining-customers`. Accessed: 2015-01-20.

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

D'Innocenzio, A. 2012. J.C. Penney gets rid of hundreds of sales. *Wall Street Journal*, 25 Jan. 2012.

Dong, J., P. Feldman, G. B. Yom-Tov. 2015. Service systems with slowdowns: Potential failures and proposed solutions. *Oper. Res.* **63**(2) 305–324.

Dwyer, J. 2010. A clothing clearance where more than just the prices have been slashed. *New York Times*, 5 Jan. 2010.

Edelson, N. M., D. K. Hilderbrand. 1975. Congestion tolls for poisson queuing processes. *Econometrica* 81–92.

Feldman, Z, A. Mandelbaum, W. A. Massey, W. Whitt. 2008. Staffing of time-varying queues to achieve time-stable performance. *Management Sci.* **54**(2) 324–338.

Fisher, M., A. Raman. 1996. Reducing the cost of demand uncertainty through accurate response to early sales. *Oper. Res.* **44**(1) 87–99.

Fortune. 2014. America's painful doctor shortage is threatening health care reform. `http://fortune.com/2014/03/04/americas-painful-doctor-shortage-is-threatening-health-care-reform/`.

Friedman, E. J. 2004. Genericity and congestion control in selfish routing. *In Proceedings of the 43rd Annual IEEE Conference on Decision and Control (CDC)*. 4667–4672.

Gans, N., G. Koole, A. Mandelbaum. 2003. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing Service Oper. Management* **5**(2) 79–141.

Gilboa-Freedman, G., R. Hassin, Y. Kerner. 2014. The price of anarchy in the markovian single server queue. *IEEE Trans. Autom. Control* **59**(2) 455–459.

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

Goyal, M., S. Netessine. 2007. Strategic technology choice and capacity investment under demand uncertainty. *Management Sci.* **53**(2) 192–207.

Goyal, M., S. Netessine. 2011. Volume flexibility, product flexibility, or both: The role of demand correlation and product substitution. *Manufacturing Service Oper. Management* **13**(2) 180–193.

Green, L., P. Kolesar. 1991. The pointwise stationary approximation for queues with nonstationary arrivals. *Management Sci.* **37**(1) 84–97.

Green, L., P. Kolesar, J. Soares. 2001. Improving the sipp approach for staffing service systems that have cyclic demands. *Oper. Res.* **49**(4) 549–564.

Green, L., P. Kolesar, W. Whitt. 2007. Coping with time-varying demand when setting staffing requirements for a service system. *Prod. and Oper. Management* **16**(1) 13.

Guo, P., P. Zipkin. 2007. Analysis and comparison of queues with different levels of delay information. *Management Sci.* **53**(6) 962–970.

Guo, P., P. Zipkin. 2009. The effects of the availability of waiting-time information on a balking queue. *Eur. J. Oper. Res.* **198**(1) 199–209.

Gurvich, I., M. A. Lariviere, A. Moreno. 2014. Staffing service systems when capacity has a mind of its own. *Working Paper*, Northwestern University.

Hassin, R. 1986. Consumer information in markets with random product quality: The case of queues and balking. *Econometrica* **54**(5) 1185–1195.

Hassin, R. 2007. Information and uncertainty in a queueing system. *Probability in the Engineering and Informational Sciences* **21**(03) 361–380.

Hassin, R., M. Haviv. 2003. *To Queue or Not to Queue: Equilibrium Behavior in Queueing Systems*. International Series in Operations Research & Management Science, Springer US.

Hassin, R., Y. Kleiner. 2010. Equilibrium and optimal arrival patterns to a server with opening and closing times. *IIE Trans.* **43**(3) 164–175.

Hassin, R., R. Roet-Green. 2013. Equilibrium in a two dimensional queueing game: When inspecting the queue is costly. Working paper, Tel Aviv University, Israel.

Haviv, M., T. Roughgarden. 2007. The price of anarchy in an exponential multi-server. *Oper. Res. Lett.* **35**(4) 421–426.

Honnappa, H., R. Jain. 2015. Strategic arrivals into queueing networks: The network concert queueing game. *Oper. Res.* **63**(1) 247–259.

Huang, T., G. Allon, A. Bassamboo. 2013. Bounded rationality in service systems. *Manufacturing Service Oper. Management* **15**(2) 263–279.

Hviid, M. 1991. Capacity constrained duopolies, uncertain demand and non-existence of pure strategy equilibria. *Eur. J. Political Econom.* **7** 183–190.

Ibrahim, R., P. L'Ecuyer. 2013. Forecasting call center arrivals: Fixed-effects, mixed-effects, and bivariate models. *Manufacturing Service Oper. Management* **15**(1) 72–85.

Jain, R., S. Juneja, N. Shimkin. 2011. The concert queueing game: to wait or to be late. *Discret. Event Dyn. Syst.* **21**(1) 103–138.

Jennings, O. B., A. Mandelbaum, W. A. Massey, W. Whitt. 1996. Server staffing to meet time-varying demand. *Management Sci.* **42**(10) 1383–1394.

Johari, R., G. Y. Weintraub, B. Van Roy. 2010. Investment and market structure in industries with congestion. *Oper. Res.* **58**(5) 1303–1317.

Juneja, S., R. Jain. 2009. The concert/cafeteria queueing problem: A game of arrivals. *Proc. Fourth Internat. ICST Conf. Performance Evaluation Methodologies and Tools* (ISCT, Gent, Belgium). 1–6.

Kalai, E., M. I. Kamien, M. Rubinovitch. 1992. Optimal service speeds in a competitive environment. *Management Sci.* **38**(8) pp. 1154–1163.

Kirman, A. P., M. J. Sobel. 1974. Dynamic oligopoly with inventories. *Econometrica* **42**(2) 279–287.

Kreps, D. M., J. Scheinkman. 1983. Quantity precommitment and bertrand competition yield cournot outcomes. *Bell J. Econom.* **14**(2) 326–337.

Lariviere, M. A., J. A. Van Mieghem. 2004. Strategically Seeking Service: How Competition Can Generate Poisson Arrivals. *Manufacturing Service Oper. Management* **6**(1) 23–40.

Levhari, D, I. Luski. 1978. Duopoly pricing and waiting lines. *Eur. Econom. Rev.* **11**(1) 17 – 35.

Li, Q., A. Ha. 2008. Reactive capacity and inventory competition under demand substitution. *IIE Trans.* **40**(8) 707–717.

Lin, Y.-T., A. Parlaktürk. 2012. Quick response under competition. *Prod. and Oper. Management* **21**(3) 518–533.

Liu, Q., D. Zhang. 2013. Dynamic pricing competition with strategic customers under vertical product differentiation. *Management Sci.* **59**(1) 84–101.

Luo, Z.-Q., J.-S. Pang, D. Ralph. 1996. *Mathematical Programs with Equilibrium Constraints*. Cambridge University Press.

Luski, I. 1976. On partial equilibrium in a queuing system with two servers. *Rev. Econom. Stud.* **43**(3) pp. 519–525.

Maggi, G. 1996. Strategic trade policies with endogenous mode of competition. *Amer. Econom. Rev.* **86**(1) 237–58.

Mandelbaum, A., W. A. Massey, M. I. Reiman, A. Stolyar, B. Rider. 2002. Queue lengths and waiting times for multiserver queues with abandonment and retrials. *Telecommunication Systems* **21**(2-4) 149–171.

Mandelbaum, A., N. Shimkin. 2000. A model for rational abandonments from invisible queues. *Queueing Systems* **36**(1-3) 141–173.

McGuire, T. W., R. Staelin. 1983. An industry equilibrium analysis of downstream vertical integration. *Marketing Sci.* **2**(2) 161–191.

Mendelson, H. 1985. Pricing computer services: queueing effects. *Communications of the ACM* **28**(3) 312–321.

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

Naor, P. 1969. The Regulation of Queue Size by Levying Tolls. *Econometrica* **37**(1) 15–24.

Netessine, S., S. Rudi, Y. Wang. 2006. Inventory competition and incentives to backorder. *IIE Trans.* **38**(11) 883–902.

Olsen, T. L., R. P. Parker. 2008. Inventory management under market size dynamics. *Management Sci.* **54**(10) 1805–1821.

Oulton, J. A. 2006. The global nursing shortage: an overview of issues and actions. *Policy, Politics, & Nursing Practice* **7**(3 suppl) 34S–39S.

Pelleau, M., L. M. Rousseau, P. L'Ecuyer, W. Zegal, L. Delorme. 2014. Scheduling agents using forecast call arrivals at hydro-québecs call centers. B. O'Sullivan, ed., *Principles and Practice of Constraint Programming*, *Lecture Notes in Computer Science*, vol. 8656. Springer International Publishing, 862–869.

Petruzzi, N., M. Dada. 2011. Newsvendor models. In *Wiley Encyclopedia of Operations Research and Management Science*. John Wiley & Sons, Inc.

Plambeck, E. L., Q. Wang. 2013. Implications of hyperbolic discounting for optimal pricing and scheduling of unpleasant services that generate future benefits. *Management Sci.* **59**(8) 1927–1946.

Porteus, E. L., H. Shin, T. I. Tunca. 2010. Feasting on leftovers: Strategic use of shortages in price competition among differentiated products. *Manufacturing Service Oper. Management* **12**(1) 140–161.

Rapoport, A., W. E. Stein, J. E. Parco, D. A. Seale. 2004. Equilibrium play in single-server queues with endogenously determined arrival times. *J. Econ. Behav. Organ.* **55**(1) 67–91.

Reynolds, S., B. Wilson. 2000. Bertrand-Edgeworth competition, demand uncertainty, and asymmetric outcomes. *J. Econom. Theory* **92**(1) 122–141.

Röller, L.-H., M. M. Tombak. 1993. Competition and investment in flexible technologies. *Management Sci.* **39**(1) 107–114.

Roughgarden, T. 2005. *Selfish Routing and the Price of Anarchy*. The MIT Press.

Roughgarden, T., É. Tardos. 2002. How bad is selfish routing? *J. ACM* **49**(2) 236–259.

Seale, D. A., J. E. Parco, W. E. Stein, A. Rapoport. 2005. Joining a Queue or Staying Out: Effects of Information Structure and Service Time on Arrival and Staying Out Decisions. *Exp. Econ.* **8**(2) 117–144.

Shaked, M., J.G. Shanthikumar. 2007. *Stochastic Orders*. Springer Series in Statistics, Springer.

Shimkin, N., A. Mandelbaum. 2004. Rational abandonment from tele-queues: Nonlinear waiting costs with heterogeneous preferences. *Queueing Systems* **47**(1-2) 117–146.

Singh, N., X. Vives. 1984. Price and quantity competition in a differentiated duopoly. *RAND J. Econom.* **15**(4) 546–554.

Small, K. A. 2015. The bottleneck model: An assessment and interpretation. *Economics of Transportation* **4**(1C2) 110 – 117.

Staelin, R. 2008. Commentary-an industry equilibrium analysis of downstream vertical integration: Twenty-five years later. *Marketing Sci.* **27**(1) 111–114.

Stein, W. E., A. Rapoport, D. A. Seale, H. Zhang, R. Zwick. 2007. Batch queues with choice of arrivals: Equilibrium analysis and experimental study. *Game & Econ. Behav.* **59**(2) 345–363.

Stigler, G. 1939. Production and distribution in the short run. *J. Political Econom.* **47** 305–327.

The Globe and Mail. 2013. Faced with doctor shortages, some emergency rooms struggle to stay open. `http://www.theglobeandmail.com/news/national/emergency-rooms-struggle-to-stay-open/article15699341/`.

Tirole, J. 1998. *The Theory of Industrial Organization*. MIT Press, Cambridge, MA.

Van Mieghem, J. A., M. Dada. 1999. Price versus production postponement: Capacity and competition. *Management Sci.* **45**(12) 1631–1649.

Veeraraghavan, S., L. G. Debo. 2009. Joining longer queues: Information externalities in queue choice. *Manufacturing Service Oper. Management* **11**(4) 543–562.

Veeraraghavan, S., L. G. Debo. 2011. Herding in queues with waiting costs: Rationality and regret. *Manufacturing Service Oper. Management* **13**(3) 329–346.

Vickrey, W. S. 1969. Congestion Theory and Transport Investment. *Amer. Econom. Rev.* **59**(2) 251–60.

Vives, X. 1989. Technological competition, uncertainty, and oligopoly. *J. Econom. Theory* **48**(2) 386–415.

Vives, X. 2001. *Oligopoly Pricing: Old Ideas and New Tools*. MIT Press Books, The MIT Press.

Wardrop, J. G., J. I. Whitehead. 1952. Correspondence. some theoretical aspects of road traffic research. *ICE Proceedings: Engineering Divisions* **1** 767–768(1).

Whitt, W. 1999. Improving service by informing customers about anticipated delays. *Management Sci.* **45**(2) 192–207.

Wu, X., F. Zhang. 2014. Home or overseas? an analysis of sourcing strategies under competition. *Management Sci.* Forthcoming.

Zhao, X., D. R. Atkins. 2008. Newsvendors under simultaneous price and inventory competition. *Manufacturing Service Oper. Management* **10**(3) 539–546.