NORTHWESTERN UNIVERSITY

Economics of Service Operations: Information, Simplified Controls and Omnichannel Services

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Operations Management

By

Abhishek Ghosh

EVANSTON, ILLINOIS

September 2021

ABSTRACT

Economics of Service Operations: Information, Simplified Controls and Omnichannel Services

Abhishek Ghosh

In this dissertation we consider how simple operational levers affect a firm's revenue and consumer surplus. In particular, we focus on information disclosure as an useful control for omnichannel services.

In the first chapter we consider a revenue-maximizing service firm that caters to price and delay-sensitive customers. The firm offers a menu of service grades where each grade is associated with a posted price and expected delay. An optimal menu size could be as large as the number of customer classes. However, in practice, we do observe that firms offer a handful number of service grades. We study the revenue loss when the firm offers a simplified menu with a few service grades. Our analysis utilizes a large system approximations under the assumption that the firm has ample capacity to serve the entire market. We set up an optimization model and make use of Taylor series and asymptotic arguments to obtain the revenue loss. We show that, under a simplified menu, the firm could lose a significant fraction of its revenue in the worst case scenario. This happens when there is significant heterogeneity between the customer classes in terms of

their delay sensitivities and their valuation for service. In contrast, noting that customer heterogeneity may typically be less extreme, we show that the firm can in fact provide a simplified menu while providing a guarantee on worst case revenue that can be obtained as a fraction of the optimal. We characterize the worst case optimal menu and provide asymptotic bounds to the worst case revenue loss as the number of customer types grow without bound. Characterization of the firm's worst case revenue loss in terms of a measure of heterogeneity can be used to guide decision making when offering a simplified menu of service grades.

In the second chapter we examine the role of information disclosure in omnichannel services. With evolving mobile technologies, an increasing number of firms are running multiple channels to serve customers. Due to the novelty of these systems, questions related to the design of such omnichannel systems and their implications for the firm and customers remain open. In particular, the question of whether or not a firm should disclose queue information to its customers in an omnichannel setting has not been extensively addressed in prior literature. Using a queuing game-theoretic framework, we address some of these open questions of design of omnichannel service system, especially focusing on the issue of congestion information disclosure and its impact on customer channel choice behavior. We benchmark the omnichannel model against a conventional single channel model, and compare these settings in terms of the firm's throughput and average consumer surplus. We find that from the firm's perspective there is no silver bullet; no channel arrangement delivers the highest throughput for all system parameters. From the customers' perspective, we once again find that neither the omnichannel nor the single channel system dominates the other in terms of the average consumer surplus for both type

of customers combined. The overall consumer surplus depends on the relative proportion of app users and non-app users in the system. Indeed, it is possible that both segments are worse-off when online ordering is offered.

In the third chapter, we extend the omnichannel setting to a competitive environment. Increasingly many firms in the quick service industry are offering digital ordering apps to customers. While the option of app-ordering is attractive to customers, still, not all firms offer an app. Even if we ignore the upfront cost of implementation of an app, it is not clear whether offering an app necessarily leads to an increase in revenue for the firm in a competitive setting. A proper evaluation needs to take into consideration the relative capacity of the firms and the sizes of their customer bases. To this end, we examine what is the best-response for a firm when faced with a competitor who offers an app. We find that it might not always be in the firm's best interest to match its competitor in offering an app.

Acknowledgements

I am indebted to my advisors Prof. Achal Bassamboo and Prof. Martin Lariviere for their constant support, guidance and mentorship throughout my doctoral studies. Without them, completing my dissertation would not have been possible. I would also like to thank Prof. Ramandeep Randhawa for helping me write my first paper, and for being patient and supportive when I was a fledgling graduate student. My special thanks to Prof. Robert Bray for his invaluable comments, suggestions and always sharing his unvarnished opinion. I would also like to thank the entire Operations Faculty at Kellogg for their support and for providing intellectual stimulation, as well as my fellow PhD students and friends for making this experience memorable.

This journey would not have been possible without motivation, love and care from my wife, Rita. I cannot thank her enough for always being there for me and believing in me even during those tough times when I had trouble believing in myself, and for being my best friend. I am also thankful to my new family, Jon, Iris and Richard, for being my home away from home.

Finally, nothing that I say would be enough to express my love and gratitude for my parents who are, and always will be, my greatest inspiration and pillars of strength. I cannot imagine my life without their fierce love for me, as well as their unwavering faith and belief in me. I thank them for always inspiring me to be a better person.

Preface

The goal of this dissertation is to consider various simplified operational controls and how they affect the economics of service systems. A primary focus of this study is to consider information disclosure in terms of what information about the service system is available to customers and in terms of the firm's lack of private information about customers. We examine how information disclosure plays a role as an operational lever for the firm, and how it plays a role in determining the type of simplified control that is available to the firm.

This dissertation consists of three chapters. The first considers a revenue-maximizing firm that cannot observe private information about customers. As a result, the firm offers an incentive-compatible menu of price and delay-differentiated service grades. Customers then self-select into their preferred service grade. In order to limit the complexity of the offered menu, the firm restricts the number of service grades it offers. We study the effect of offering a simplified menu of service grades on the firm's revenue.

The second chapter deals with omnichannel services where customers can order remotely using a digital app. In particular, this chapter examines whether or not an omnichannel firm should disclose congestion information about the service system to customers via the app, and further examines how this decision affects customers' ordering

strategy. In this case, the decision of whether or not to disclose congestion information acts as an operational lever for the firm. We evaluate how the firm's revenue, and consumer surplus is affected by this control.

Finally, the third chapter considers competing service providers, with the goal of examining whether offering an app always leads to an increase in the firm's revenue when its competitor firm offers an app. An app reveals congestion information to customers, and also offers customers the ability to order whenever and from wherever they want. Thus, in a competitive setting, a firm's decision of whether or not to offer an app potentially affects how customers choose from which firm to seek service. We examine the impact of this decision of whether or not to offer an app on the firm's revenue when it faces a competitor who offers an app.

Dedication

To my family

Table of Contents

List of Tables

List of Figures

CHAPTER 1

Value of Simple Menus with Price and Delay Sensitive Customers (Joint work with Achal Bassamboo and Ramandeep Randhawa)

1.1. Introduction

Motivation and Research Question. Firms serving price and time-sensitive customers often provide a menu of differentiated service grades with posted prices and lead-times (or delays) where the impatient customers are charged a premium in return for expedited service. For example, firms like Amazon have multiple delivery options where patient customers can opt for a free regular delivery whereas impatient customers can pay a delivery fee for same-day or expedited delivery. Parcel delivery services like UPS offer expedited overnight shipping for a higher price as compared to regular ground shipping. Most theme park operators provide their customers with the option of paying for a pass (for example, Universal offers Express Pass and Six Flags offers *THE FLASH Pass*) which essentially acts as priority access and allows its customers to skip the line and reduce the time they spend waiting. We can refer to numerous other similar examples in the areas of communication, transportation and government services, where the customers are simultaneously differentiated based on their willingness to pay for service and their sensitivity to delays experienced in service. Revenue maximizing firms use this differential pricing as a tool to extract higher revenue from the less patient customer base.

Customers are heterogeneous and have *private information* about their own preferences relating to their willingness to pay for service and their sensitivity to delays. Some customers might be willing to pay higher than the others for instant service (valuation) and some customers might value faster service more than the others (delay sensitivity). Throughout the paper, we use *customer class* to refer to customer "type" which is characterized by the valuation that the customer has for the service, and the delay sensitivity of the customer. Multiple customer classes could potentially be served by a single service grade with a posted price and lead-time. By offering a menu of price and delay differentiated service grades, the firm lets individual customers self-select into their preferred service grade. Customers are self-interested in their choice of service grades. The firm does not possess information about the individual customer's preferences. As a result, a mechanism needs to be designed so that the customers choose the grade of service that is designed for them. In this sense, the mechanism/menu must be incentive compatible.

Since the firm caters to a heterogeneous customer base, it is expected that a revenue maximizing menu should offer a large number of service grades which increases with the number of heterogeneous customer classes. However in practice, we see that in order to limit the complexity of the offered menu, the number of service grades offered by these firms is typically limited. By offering a menu with limited number of service grades, the firm could potentially leave money on the table. In our paper, we study the following research question: *How much revenue can the firm lose by offering a simplified menu that has limited number of service grades as compared with the revenue maximizing optimal menu?*

Methodology. Our analysis utilizes a large system approximation where we assume that the firm has *ample capacity* to serve the entire market. This approximation implies that the size of each customer class proportionately *scales* with firm's capacity while maintaining the stability of the system. In this setting of ample capacity, we note that any work conserving policy would result in zero lead times. As a result, it would be socially optimal and incentive compatible to serve everyone with a price and lead-time of zero. However, revenue maximization would entail artificially inflating the lead times to differentiate across customer classes. This would dis-incentivize the impatient classes from joining the cheaper and slower service grades. Therefore, the delays in our framework do not arise from the congestion in the system due to queuing effects, rather they arise solely due to induced server idleness necessary for satisfying the incentive compatibility constraints. Our ample capacity assumptions allows us to focus on these delays without being mired with tackling the queueing dynamics.

In order to answer our research question, we study the firm's revenue under a menu with limited number of service grades possibly fewer than the number of customer classes, as a fraction of the maximum possible revenue under the optimal menu. Using this study, we capture the performance of the firm in terms of the revenue relative to the optimal revenue. Thus, a higher value of this ratio corresponds to a lower optimality gap and a lower value indicates significant revenue loss and sub-optimal performance by the firm. With this ratio as the objective function, we measure the performance of a simplified menu using a worst case analysis by solving a minimization problem in terms of the valuation and delay sensitivities of the customer classes. The solution of this optimization problem corresponds to the maximum possible revenue loss by the firm and hence we refer to it as

the *worst case revenue ratio* (WCRR) and the resulting valuation and delay sensitivities as the *worst case parameters*. We derive structural properties for the optimal menu under the worst case scenario and make use of Taylor series and asymptotic arguments to derive the solution to this problem.

Contribution and Summary of Results. We show that, in general, the firm could lose a significant fraction of its revenue by offering a limited menu of service grades in the worst case scenario. That is, we show that the WCRR converges to zero as the number of customer classes grows without bound. We further investigate the parameter regime in which these significant losses to revenue are realized and find that this happens in settings in which the heterogeneity in delay sensitivity and valuations is significant even between any two customer classes.

Noting that customer heterogeneity may typically be less extreme, we analyze the case of bounded heterogeneity between customer classes. We find that there is a big difference between the cases of unlimited and limited heterogeneity with respect to how customer classes are segmented in the optimal menu under the worst case scenario, that we refer to as the optimal worst case menu. In the unlimited heterogeneity setting, the optimal worst case menu separates out each customer class customer class is differentiated from each other. In contrast, in the limited heterogeneity setting, some customer classes may be pooled. Thus, in the latter case, the number of candidate segmentations of customer classes grows exponentially with the number of customer classes. We show that this complexity can be reduced so that one only needs to consider a strict subset of the possibilities that is quadratic in the number of customer classes. We characterize the worst case optimal menu, and use it to show that, under limited heterogeneity, the firm

can in fact provide a simplified menu while providing a guarantee on worst-case revenue that can be obtained as a fraction of the optimal. Mathematically, we prove that the WCRR for the firm can be bounded away from zero even if the number of classes grows without bound. Further, we characterize the firm's worst case revenue loss in terms of a measure of heterogeneity, which can be used to guide decision making when offering a simplified menu of service grades.

1.2. Literature Review

Our paper contributes to the vast literature on pricing and scheduling in queuing systems with strategic customers. The work on strategic customers in queues dates back to the seminal paper by Naor (1969). Mendelson and Whang (1990) was the first paper to look at the pricing problem from a social welfare maximization perspective when serving multiple customers types with type dependent delay sensitivities and linear delay cost. Subsequent papers have considered the social welfare maximization problems (Van Mieghem, 2000; Hsu et al., 2009). It turns out that the pricing mechanism is *incentive compatible* and the optimal scheduling policy is work-conserving. This same problem when studied from a revenue maximization perspective becomes fairly complicated.

Afeche (2013) addresses this problem in the revenue maximization setting in presence of two customer types and shows that externality pricing and delay cost minimization are no longer optimal in this setting. Moreover, the paper shows that in the optimal solution, the service provider artificially delays one customer type ("strategic delay") beyond what is obtained from the work-conserving policy. Yahalom et al. (2006) extends this problem to non-linear delay costs and multiple customer types. Partial extensions of this approach

to more than two customer types in an M/M/1 setting is studied in Katta and Sethuraman (2005) and Afeche and Pavlin (2016) under additional assumptions on the relationship between valuations and delay costs.

Doroudi et al. (2013) considers an M/G/1 queue where arriving customers draw valuation from a common distribution and have waiting costs that are proportional to their realized valuations. The bulk of their analysis focuses on offering a continuum of priorities but they do demonstrate numerically that a coarse priority scheme with a limited number of priority classes performs very well. They explicitly compute the optimal menu of prices in closed form for some specific customer valuation distributions. Katta and Sethuraman (2005) considers a model with multiple, finite customer types with each customer type having a constant type-dependent valuation and delay sensitivity coefficient. Under assumptions on valuation and delay coefficients, the authors characterize the structure of the optimal pricing and scheduling policy. Furthermore, similar to our setting, they consider the case where the service provider is restricted to use a limited number of service levels. Although, these papers focus on characterizing the optimal policy, due to the complex nature of the problem, there is a lack of insight into the value of offering simplified menus. Nazerzadeh and Randhawa (2018) use an asymptotic analysis to show that, when delay sensitivities are linear or sub-linear in the customer valuation, a very coarse priority scheme is sufficient; in a large system, two levels of priority is asymptotically optimal and capture nearly all of the possible system value.

Our work focuses on the revenue optimization problem and best relates with Nazerzadeh and Randhawa (2018), Afeche and Pavlin (2016), Doroudi et al. (2013) and Katta and Sethuraman (2005). In relation to these papers, our paper is similar in the

sense that, we seek to understand the value of offering a simplified menu. While most of these papers focus on the exact characterization of the optimal policy, we take a large system asymptotic approach similar to Nazerzadeh and Randhawa (2018) and Maglaras et al. (2013). To study our problem, we assume that the firm has *ample capacity*, which gets rid of queuing related congestion. Consequently, we do not focus on the scheduling problem. We perform a worst-case analysis of the optimality gap in the firm's revenue under limited offering. We consider discrete customer types, similar to Katta and Sethuraman (2005), and present the asymptotic lower bound on the firm's revenue loss as the number of customer types grow without bound. We make use of Taylor series arguments for our analysis, similar to Nazerzadeh and Randhawa (2018) and Maglaras and Zeevi (2003).

In our paper, we study the worst case scenario for the firm. As an outcome of this worst case analysis, we find that the delay sensitivities would be increasing in valuations, i.e., higher valuation customers will be more delay sensitive. In contrast, previously mentioned papers study the value of offering a simplified menu under the assumption of a monotone relationship between customer valuations and the delay cost. For e.g., Afeche and Pavlin (2016) considers the case in which the customer delay sensitivity coefficient is affine in valuation whereas both Katta and Sethuraman (2005) and Nazerzadeh and Randhawa (2018) assume that delay sensitivities are sublinear in valuations. Furthermore, depending on the waiting cost structure, the revenue maximizer may pool some types together, impose a common price and offer the same expected wait, see Katta and Sethuraman (2005). Additionally, a revenue maximizer may use a complex service discipline that may pool customers or exclude some or impose strategic delay, see Afeche and Pavlin (2016).

Similarly, we find in our worst-case analysis, that the optimal menu might pool multiple classes into a single service grade or exclude some of the customer classes or offer price-delay differentiation through strategic delay. Finally, we focus on static price-delay menus in contrast to dynamic price-lead time quotations which is studied in Plambeck (2004), Çelik and Maglaras (2008) and Ata and Olsen (2013). Our focus in this paper is on posted pricing; one could consider other price-service mechanism such as priority auctions, see Afèche and Mendelson (2004).

Finally, there are papers which consider similar setting but address different research questions. Gurvich et al. (2018) compares how priority scheme is implemented by a revenue maximizer and a social planner, and addresses the question of how priority schemes affect consumer surplus. Maglaras and Zeevi (2003) studies the pricing and capacity sizing problem for systems with shared resources under revenue and social optimization objectives. Maglaras et al. (2013) considers this problem in a large system asymptotic regime and uses these asymptotic finding to provide interesting contrasts between social welfare and revenue optimization.

1.3. Model

We model price and time sensitive customers who are heterogeneous in their valuations for service, and delay sensitivities. We use N to denote the total number of customer classes. All customers of class i have valuation (willingness-to-pay) v_i for service and incur a linear delay cost of h_i per unit time spent in the queue waiting for service to commence. We use λ_i to denote the arrival rate for customer class i. Without loss of generality, we index customer classes in decreasing order of their valuations, i.e. $v_1 > v_2 > \cdots > v_N$.

Thus, $(v_i, h_i, \lambda_i) \in \mathbb{R}^3_+$ fully characterizes class i of customers. We use v, h and λ to denote the valuation, delay sensitivity and arrival rate vectors $(v_1, v_2, \ldots, v_N)$, $(h_1, h_2, \ldots, h_N)$ and $(\lambda_1, \lambda_2, \ldots, \lambda_N)$ respectively. We use $\overline{\lambda}_l$ to denote $\sum_{m=1}^{l} \lambda_m$ for brevity, throughout the rest of this paper.

Upon arrival to the service system, customers are faced with a menu of K service grades that are differentiated in terms of price and delay offered by a monopolistic firm. The parameters for the k^{th} service grade are denoted by $(p_k, d_k) \in \mathbb{R}^2_+$ where p_k is the price and d_k is the delay experienced, for all $k = 1, 2, \ldots, K$. A menu with K service grades is denoted by $\{(p_k, d_k) \mid 1 \leq k \leq K\}$. Without loss of generality, we index service grades in decreasing order of the prices, i.e. $p_1 > p_2 > \cdots > p_K$. We model the utility of a customer of class i who pays price p for service and experiences a delay d, as a linear function of her valuation net of price paid and the delay cost incurred corresponding to the total time spent by her in the queue, which is given by $v_i - p - h_i d$. Customers decide whether or not to join service, and if they do, which service grade to choose so that their individual utility is maximized. Thus, for a given customer of class $i \in \{1, 2, \ldots, N\}$, defining

$$j(i) = \underset{k=1,2,\ldots,K}{\arg\max} (v_i - p_k - h_i d_k), \tag{1.1}$$

the customer joins service grade $j(i)$ if $v_i - p_{j(i)} - h_i d_{j(i)} \geq 0$, and chooses not to join service if, $v_i - p_{j(i)} - h_i d_{j(i)} < 0$. If customers are indifferent between joining any two service grades, we assume that they always join the service grade with lower delay. Based

on the customer's choice of service, the effective arrival rate for service grade k is,

$$\phi_k(p_k, d_k) = \sum_{i \in \mathcal{I}_k} \lambda_i$$

where $\mathcal{I}_k$ is the set of all customer classes joining service grade k.

The firm's problem is to offer a menu of K service grades such that it's revenue is maximized. We make the following relaxation regarding the firm's capacity:

Assumption 1. *We assume that the firm can offer any delay for all classes,* $d_k \geq 0$ *for all* $k \in \{1, 2, \ldots, K\}$.

Assumption 1 provides analytical tractability. Further, Assumption 1 is a good approximation for *large scale* queuing systems (see §4.1 in Maglaras et al., 2013) where the size of each customer class proportionately *scales* with capacity keeping the system stable by maintaining an overall constant throughput of strictly less than one, holding all other parameters of the problem fixed. In this way, the customer population grows large, but the characteristics and behavior of individual customers remain the same. Therefore, any non-zero delay offered by the firm in our setting can be viewed as "strategic delay" (see Afeche, 2013) which implies that the firm may strategically induce server idleness in order to satisfy incentive compatibility. In particular, strategic delay allows the firm to offer differentiated service grades and charge higher valuation customers a premium for service grades with lower delays.

The firm's optimization problem for the best simplified menu with K grades of service is given by,

$$\pi_N^K(v,h) = \sup_{p_k,d_k} \sum_{k=1}^{K} \phi_k(p_k,d_k)p_k \quad \text{s.t. } p_k \geq 0,\ d_k \geq 0 \text{ for all } k = 1,2,\ldots,K. \tag{1.2}$$

In contrast, for the optimal menu that maximizes the revenue, depending on the valuation and the delay sensitivities of the customers, the firm may need to offer up to N service grades. We use $\pi_N^*(v,h)$ to denote the revenue of this optimal menu.

Our goal is to study the loss of revenue for a firm offering K service grades, where $K \leq N$, in comparison to the revenue for the optimal menu under the worst case valuation and the delay sensitivities. To this end, we are interested in solving the following optimization problem which represents our worst case analysis:

$$R_N^K = \inf_{v>0,h>0} \left\{ \frac{\pi_N^K}{\pi_N^*} \right\}. \tag{1.3}$$

We refer to R_N^K as the *worst case revenue ratio* (WCRR). We refer to the optimal menu, which generates a revenue of π_N^* under the optimal solution to (1.3), as the *worst case optimal menu* (WCOM). Thus, R_N^K is the fraction of the optimal revenue, π_N^*, that the firm generates by offering a simplified menu with K service grades, under the worst case valuations and delay sensitivities. A higher value of R_N^K would indicate that a simplified menu is valuable whereas a lower value would indicate that the firm may lose a lot of revenue by offering a simplified menu.

1.4. Analysis: Fundamentals

In this section, we define a set of menus of service grades, and their properties, which will be essential for the rest of our analysis. In particular, we use these definitions to characterize the worst case optimal menu, WCOM (see the definition after optimization problem (1.3)), for our worst case analysis for a given K. We begin by defining the following customer segmentation.

Definition 1. (Customer Segmentation σ) *We say a menu of service grades,* $\{(p_l, d_l) \mid 1 \leq l \leq L\}$*, induces a customer segmentation* $\sigma(i_1, i_2, \ldots, i_L)$*, where* $0 = i_0 < i_1 < \cdots < i_L \leq N$*, if customer classes* $i_{l-1}+1, i_{l-1}+2, \ldots, i_l$ *join service grade* l *for all* $l = 1, 2, \ldots, L$*, and customer classes* $i_L + 1, i_L + 2, \ldots, N$ *do not join service, where* L *denotes the number of service grades.*

Customer segmentation σ implies that blocks of consecutive customer classes join consecutive service grades. Next, we define a set of menus that induces customer segmentation σ.

Definition 2. ($\sigma-$Inducing Menus) *We define* $\mathbb{M}_L$ *as the set of all menus that offers* L *service grades,* $\{(p_l, d_l) | l = 1, 2, \ldots, L\}$*, such that these menus induce some customer segmentation* $\sigma(i_1, i_2, \ldots, i_L)$*, and the utility of customer class* i_l *resulting from joining service grade* (p_l, d_l) *is zero, i.e.,* $v_{i_l} - p_l - h_{i_l} d_l = 0$ *for all* $l = 1, 2, \ldots, L$*. We define* $\mathbb{M} = \cup_{L=1}^{N} \mathbb{M}_L$.

Now that we have defined customer segmentation σ and a set of menus that induces σ, we present the following lemma, which uses these definitions to establish the characteristics of the WCOM.

Lemma 1. *The optimal menu, WCOM, in the worst case analysis as defined in (1.3), belongs to set* $\mathbb{M}$*, when the firm offers a single service grade, i.e.* $K = 1$.

The prices and delays offered on revenue-maximizing menus in set $\mathbb{M}$ are characterized by the proposition below. In addition, we present a condition, which involves the delay sensitivities and the arrival rates of the customer classes. This condition is necessary in order for the menus of set $\mathbb{M}$ to be able to delay-differentiate customer classes.

Proposition 1. *Consider two service grades* (p_j, d_j) *and* (p_k, d_k) *on a revenue - maximizing menu in set* $\mathbb{M}$*, where* $j < k$*, and for corresponding customer classes* i_j *and* i_k *the following holds:*

(i) The delay sensitivities satisfy the relation $\dfrac{h_{i_j}}{h_{i_k}} > \dfrac{\overline{\lambda}_{i_k}}{\overline{\lambda}_{i_j}}$ *where* $\overline{\lambda}_k = \sum_{i=1}^{k} \lambda_i$.

(ii) The prices satisfy $p_k = v_{i_k} - h_{i_k} d_k$ *and the delays satisfy*

$$d_k = \begin{cases} 0 & k = 1, \\ \frac{v_{i_{k-1}} - v_{i_k}}{h_{i_{k-1}} - h_{i_k}} & k > 1. \end{cases}$$

The benefit of delay differentiating any two customer classes comes from the fact that the customer class with higher valuation is charged a higher price for a faster service. Simultaneously the lower valuation customer class is charged a lower price for a service that is sufficiently "degraded," by offering more delay, to disincentivize the higher valuation customer class from joining it. The offered delay is smaller if these two customer classes are relatively more delay sensitive to one another. On the other hand, as the customer classes become less delay sensitive relative to one another, the offered delay that is required to differentiate them becomes larger. Thus, the revenue coming from the

higher priced service grade should compensate for the reduction in revenue coming from the lower priced service grade in order for delay-differentiation to be beneficial to the firm. Proposition 1(i) highlights this trade-off and says that if customer classes i_j and i_k are delay-differentiated, then it implies that these two customer classes are adequately delay sensitive relative to one another. This would ensure that the resultant price reduction for the slower service grade, by offering higher delay, does not outweigh the benefit of offering delay differentiation. This result plays an important role in determining the structure of the solution to our worst case analysis, especially in §1.7, where we analyze the case with limited heterogeneity.

In §3.1 we present the revenue-maximizing menu that is optimal for the firm to offer, when there are two customer classes. We discuss how this menu changes as the valuation and delay sensitivities of the customer classes change. In addition, we provide the worst case analysis for the firm's revenue when it offers a single service grade in the presence of two customer classes. As the number of classes grow beyond two, characterizing the revenue-maximizing menu becomes cumbersome, and hence in §3.2, we analyze the WCRR when the firm faces more than two customer classes. We show that, the worst case is realized as the consecutive customer classes become infinitely more delay sensitive and have infinitely more valuation. Finally, in §3.3 we formalize the idea of *limited heterogeneity*, which refers to the valuations and delay sensitivities of the customer classes being bounded, and present the worst case analysis for this case.

1.5. Analysis: Two Customer Classes ($N = 2$)

In this section we present the revenue-maximizing menu for the firm in the presence of two customer classes, i.e. $N = 2$. In the subsequent section, we use this menu to analyze (1.3), when the firm offers a single service grade, i.e. $K = 1$. Afeche (2013) and Maglaras et al. (2013) provides the optimal price/lead-time menu for two customer classes for a capacity constrained firm. As a result, some of our results in §3.1.1 would overlap with theirs.

1.5.1. Properties of Optimal Menu

Consider two customer classes with valuations and delay sensitivities (v_1, h_1) and (v_2, h_2) respectively. The optimal menu could offer either one or two service grades catering to the two customer classes. We solve for the optimal price-delay menu, $\{(p_k, d_k) \mid 1 \leq k \leq 2\}$, and derive conditions in terms of the customer primitives, under which it is optimal to offer this menu. Solution to the following optimization problem yields the revenue-maximizing menu when there are two customer classes:

$$
\begin{aligned}
\max_{p_1,d_1,p_2,d_2} \quad & \lambda_1 p_1 + \lambda_2 p_2 \\
\text{s.t.} \quad & v_1 - p_1 - h_1 d_1 \geq 0, && \text{(IR1)} \\
& v_2 - p_2 - h_2 d_2 \geq 0, && \text{(IR2)} \\
& v_1 - p_1 - h_1 d_1 \geq v_1 - p_2 - h_1 d_2, && \text{(IC1)} \\
& v_2 - p_2 - h_2 d_2 \geq v_2 - p_1 - h_2 d_1, && \text{(IC2)} \\
& p_1 \geq 0,\ p_2 \geq 0,\ d_1 \geq 0,\ d_2 \geq 0.
\end{aligned}
\tag{1.4}
$$

Participation constraints (IR1) and (IR2) ensure that the customer classes get non-negative utility from joining service. *Incentive compatibility* constraints (IC1) and (IC2) ensure that the two classes join their respective service grades. The following lemma presents the solution to problem (1.4), and hence completely characterizes the revenue-maximizing menu in terms of the customer primitives:

Lemma 2. *If* $\left(\frac{h_1}{h_2} > 1 + \frac{\lambda_2}{\lambda_1}\right)$ *then it is optimal to offer the following menu:*

i) If $\left(\frac{v_1}{h_1} < \frac{v_2}{h_2}\right)$ *then optimal number of service grades,* $K^* = 2$*, and* $p_1 = v_1$*,* $d_1 = 0$*,* $p_2 = v_2 - h_2\left(\frac{v_1 - v_2}{h_1 - h_2}\right)$*,* $d_2 = \frac{v_1 - v_2}{h_1 - h_2}$*. Optimal revenue for the firm is* $\lambda_1 v_1 + \lambda_2\left(v_2 - h_2\left(\frac{v_1 - v_2}{h_1 - h_2}\right)\right)$*.*

ii) If $\left(\frac{v_1}{h_1} \geq \frac{v_2}{h_2}\right)$ *then* $K^* = 1$*, and* $p_1 = v_1$*,* $d_1 = 0$*. Optimal revenue for the firm is* $\lambda_1 v_1$*.*

If $\left(\frac{h_1}{h_2} \leq 1 + \frac{\lambda_2}{\lambda_1}\right)$ *then it is optimal to offer the following menu:*

i) If $\left(\frac{v_1}{v_2} < 1 + \frac{\lambda_2}{\lambda_1}\right)$ *then* $K^* = 1$*, and* $p_1 = v_2$*,* $d_1 = 0$*. Optimal revenue for the firm is* $(\lambda_1 + \lambda_2)v_2$*.*

ii) If $\left(\frac{v_1}{v_2} \geq 1 + \frac{\lambda_2}{\lambda_1}\right)$ *then* $K^* = 1$*, and* $p_1 = v_1$*,* $d_1 = 0$*. Optimal revenue for the firm is* $\lambda_1 v_1$*.*

We note that, for the revenue-maximizing menu to provide delay differentiated service grades, it is necessary that the higher valuation customers also have higher delay sensitivity, i.e. if $v_i > v_j$ then $h_i > h_j$. Moreover, for delay differentiation to be optimal, it is necessary that delay sensitivity per unit valuation for class 1 be greater than that of

class 2, i.e. $\frac{h_1}{v_1} > \frac{h_2}{v_2}$. This ensures that the firm can offer a delay differentiated service grade to customer class 2 with a non-negative price, $p_2 = v_2 - h_2\left(\frac{v_1 - v_2}{h_1 - h_2}\right)$, otherwise, if $\frac{h_1}{v_1} \leq \frac{h_2}{v_2}$ then offering a single grade becomes optimal (This is same as the *price condition* in Proposition 4 and §6.3 of Afeche (2013) for the case with *ample capacity*). We observe that there is a positive delay in the second service grade, i.e. $d_2 > 0$, when delay differentiation is optimal. This is known as *strategic delay* in the literature. Strategic delay ensures incentive compatibility of the mechanism, i.e. the higher valuation class 1 joins the service with higher price p_1 and lower delay d_1, as compared to, the "degraded" service grade 2, with price $p_2 < p_1$ and delay $d_2 > d_1$. Furthermore, we note that the condition $\left(\frac{h_1}{h_2} \geq 1 + \frac{\lambda_2}{\lambda_1}\right)$ in Lemma 2, which is necessary for delay-differentiation to be optimal, can also be obtained by applying Proposition 1. This is same as the *segment-size condition* in §6.3 of Afeche (2013).

1.5.2. Worst Case Analysis

Now that we have characterized the properties of the optimal menu, we use it to solve (1.3), and study the worst case revenue loss associated with offering a simplified menu. The worst case scenario would be realized when it is optimal for the firm to offer two service grades and the firm offers a single service grade. To see this, we note that if it were optimal to offer a single service grade, then the firm would be acting optimally by offering a single service grade, and hence this couldn't be the worst case. Thus, we are interested in the following optimization problem, which solves for the WCRR:

$$R_2^1 = \inf_{v>0, h>0} \left\{ \frac{\pi_2^1}{\pi_2^*} \right\}.$$

The service grade offered by the firm would induce one of two possible customer segmentations. Either both customer classes would be *pooled* into the same service grade or only the higher valuation customer class 1 would join service. Given the customer primitives, this leads to two possible revenue-maximizing service grades that the firm could offer; either $(p_1 = v_1, d_1 = 0)$ or $(p_2 = v_2, d_2 = 0)$. Thus, the firm's revenue is $\pi_2^1 = \max\{\lambda_1 v_1, (\lambda_1 + \lambda_2)v_2\}$. Applying Lemma 2, we note that the optimal menu consists of the following two service grades,

$$(p_1 = v_1, d_1 = 0) \text{ and } \left(p_2 = v_2 - h_2\left(\frac{v_1 - v_2}{h_1 - h_2}\right), d_2 = \frac{v_1 - v_2}{h_1 - h_2}\right).$$

Thus, the WCRR can be written as,

$$R_2^1 = \inf_{v>0, h>0} \left\{ \frac{\max\{\lambda_1 v_1, (\lambda_1 + \lambda_2)v_2\}}{\lambda_1 v_1 + \lambda_2\left(v_2 - h_2\left(\frac{v_1 - v_2}{h_1 - h_2}\right)\right)} \right\} \tag{1.5}$$

$$\text{s.t. } v_1 > v_2, \tag{1.6}$$

$$\frac{h_1}{h_2} > \frac{v_1}{v_2}. \tag{1.7}$$

Constraint (1.6) reflects our convention of indexing customer classes by decreasing order of their valuations. Constraint (1.7) is the price condition which ensures that price p_2 and delay d_2 are positive. The following lemma presents the solution to problem (1.5).

Lemma 3. *The WCRR for a firm offering a single service grade, i.e. $K = 1$, when there are two customer classes, i.e. $N = 2$, and the arrival rates λ_1 and λ_2 are fixed, is given by,*

$$R_2^1 = \frac{\lambda_1 + \lambda_2}{\lambda_1 + 2\lambda_2}.$$

The worst case is realized as the ratio of consecutive delay sensitivities grow without bound, i.e. $\frac{h_1}{h_2} \to \infty$. Moreover, under the optimal solution to (1.5), the valuations are such that $\frac{v_1}{v_2} = \frac{\lambda_1+\lambda_2}{\lambda_1}$, i.e. both service grades $(v_1, 0)$ and $(v_2, 0)$ would yield the same revenue. In the expression for price, $p_2 = v_2 - h_2\big(\frac{v_1-v_2}{h_1-h_2}\big)$, we can interpret the term $h_2\big(\frac{v_1-v_2}{h_1-h_2}\big)$ as the cost of offering delay-differentiation, as it represents the delay cost associated with offering differentiated service grades. Since, $\frac{h_1}{h_2} \to \infty$, the offered delay $d_2 = \frac{v_1-v_2}{h_1-h_2}$ approaches zero, and thus the cost of offering delay differentiation goes to zero as well. This implies that the two customer classes become progressively "easier" to differentiate, and hence the firm becomes progressively worse off by offering a single service grade. Thus, under the worst case valuations and delay sensitivities, the firm can guarantee $\frac{\lambda_1+\lambda_2}{\lambda_1+2\lambda_2}$ fraction of the optimal revenue by offering a single service grade.

1.6. Analysis: Multiple Customer Classes ($N > 2$)

Characterizing the revenue-maximizing menu, as we have done in §1.5.1, becomes cumbersome when there are more than two customer classes. Hence, in this section, we present the worst case analysis for the firm's revenue when there are more than two customer classes, i.e. $N > 2$, and when the firm offers fewer than N service grades. To this end, we are interested in characterizing the WCRR, R_N^K, as defined in (1.3). The theorem below summarizes the result of our worst case analysis and characterizes the WCRR.

Theorem 1. (Worst Case Revenue Ratio) *The WCRR for a firm, offering a menu of K service grades when there are N customer classes, is such that, $\frac{1}{N} \leq R_N^K \leq \frac{K}{N}$.*

In the remainder of this section, we set up optimization problem (1.3) and highlight certain characteristics of the optimal solution. This will help develop intuition and will lead into the subsequent sections. we begin with our analysis of (1.3) for $K = 1$, i.e. when the firm offers a single service grade. we obtain an exact expression for R_N^1. However, deriving an exact expression for R_N^K when $K > 1$, is cumbersome. Therefore, we use the structure of the optimal solution for $K = 1$ and derive an upper bound to (1.3) when $K > 1$, i.e. when the firm offers multiple service grades.

Single Service Grade ($K = 1$). When there are N customer classes and the firm offers a single service grade, i.e. $K = 1$, there are N possible revenue-maximizing service grades that the firm could offer. The service grade denoted by $(p_k = v_k, d_k = 0)$ would result in customer classes $1, 2, \ldots, k$ being pooled into one service grade and the remaining classes not joining service, for all $k = 1, 2, \ldots, N$. The optimal menu would offer N incentive compatible service grades targeting each of the N customer classes. Applying Lemma 1, we know that the worst case optimal menu belongs to set $\mathbb{M}$. Hence, we can apply Proposition 1 to characterize the prices and delays offered by the optimal menu. Thus, the optimal menu would offer N service grades $\{(p_k, d_k)|1 \leq k \leq K\}$, where $p_1 = v_1$, $d_1 = 0$, and $p_k = v_k - h_k\big(\frac{v_{k-1}-v_k}{h_{k-1}-h_k}\big)$ and $d_k = \frac{v_{k-1}-v_k}{h_{k-1}-h_k}$ for all $k = 2, 3, \ldots, N$. Thus, the WCRR is given by,

$$
\begin{aligned}
R_N^1 = \inf_{\lambda>0, v>0, h>0} & \left\{ \frac{\max_{1\leq k\leq N}\{\overline{\lambda}_k v_k\}}{\lambda_1 v_1 + \sum_{k=2}^{N} \lambda_k\big(v_k - h_k\big(\frac{v_{k-1}-v_k}{h_{k-1}-h_k}\big)\big)} \right\} \\
& \text{s.t. } v_k > v_{k+1},\ \frac{h_k}{h_{k+1}} > \frac{v_k}{v_{k+1}},\ \text{for all } k = 1, 2, \ldots, N-1.
\end{aligned} \tag{1.8}
$$

The worst case is realized as the ratio of the consecutive delay sensitivities grow without bound, i.e. $\frac{h_k}{h_{k+1}} \to \infty$, for all $k = 1, 2, \ldots, N-1$. As a result, any two customer classes could be differentiated by offering infinitesimally small delay. This allows for maximum possible delay-differentiation as the delay cost, $h_k\big(\frac{v_{k-1}-v_k}{h_{k-1}-h_k}\big)$, goes to zero. Thus, we can rewrite (1.8) as follows:

$$R_N^1 = \inf_{\lambda>0, v>0} \left\{ \frac{\max_{1\le k\le N}\{\overline{\lambda}_k v_k\}}{\sum_{k=1}^N \lambda_k v_k} \right\} \tag{1.9}$$
$$\text{s.t. } v_k > v_{k+1}, \text{ for all } k = 1, 2, \ldots, N-1.$$

We solve problem (1.9) exactly, and Theorem 1 establishes this result. Under the optimal solution to (1.9), the valuations are such that they satisfy $\overline{\lambda}_k v_k = \overline{\lambda}_{k+1} v_{k+1}$ for all $k = 1, 2, \ldots, N-1$. Using this relationship between the worst case valuations and arrival rates, we can reformulate (1.9) as an optimization problem over arrival rates only. In particular, by dividing both the numerator and the denominator in (1.9) by $\overline{\lambda}_k v_k$ we get,

$$R_N^1 = \inf_{\lambda>0} \left\{ \frac{1}{\sum_{k=1}^N \frac{\lambda_k}{\overline{\lambda}_k}} \right\}, \tag{1.10}$$
$$\text{where } \overline{\lambda}_k = \sum_{i=1}^{k} \lambda_i \text{ for all } k = 1, 2, \ldots, N.$$

Equivalently, we could also express (1.9) as an optimization problem over valuations only. Thus, minimizing the revenue ratio over valuations is equivalent to minimizing the revenue ratio over arrival rates. Optimal solution to problem (1.9) is achieved as the ratio of the consecutive valuations grow without bound, i.e. $\frac{v_k}{v_{k+1}} \to \infty$. This implies that $\frac{\overline{\lambda}_{k+1}}{\overline{\lambda}_k} \to \infty$, which can be shown to be equivalent to $\frac{\lambda_{k+1}}{\lambda_k} \to \infty$. Hence, the worst case is realized as

the valuations for consecutive customer classes get infinitely higher and the arrival rates get infinitely smaller in such a way that $\overline{\lambda}_k v_k$ remains constant for all $k = 1, 2, \ldots, N$.

Multiple Service Grades ($K > 1$). Now, if the firm offers a menu with more than one service grade, i.e. $K > 1$, deriving an exact expression for R_N^K would be cumbersome. Lemma 1 characterizes the worst case optimal menu for $K = 1$. However, characterizing the worst case optimal menu and the expression for the revenue associated with this menu, would not be as straight forward for $K > 1$, as compared to $K = 1$. Deriving an upper bound for R_N^K, would be sufficient for us to generate insights and analyze the WCRR. Since, we are not interested in characterizing the optimal menu for $K > 1$, we argue that, in (1.3) if we use revenue expressions π_N^K and π_N^* that correspond to feasible menus then it would result in an upper bound to R_N^K. To see this, we note that, the correct revenue expressions in (1.3), would result in the least value of the revenue ratio, resulting in the WCRR. Any other feasible revenue expression in (1.3), can only result in a higher value of the revenue ratio, since (1.3) is a minimization problem. Hence, this would result in an upper bound. Thus, we assume that both the optimal menu and the K-service grade revenue-maximizing menu offered by the firm belong to set $\mathbb{M}$ of σ-inducing menus. The K-service grade revenue-maximizing menu, $\{(p_k, d_k) | 1 \leq k \leq K\}$ which is given by Proposition 1, is such that $p_1 = v_1, d_1 = 0$, and $p_k = v_{i_k} - h_{i_k}\left(\frac{v_{i_{k-1}} - v_{i_k}}{h_{i_{k-1}} - h_{i_k}}\right)$ and $d_k = \frac{v_{i_{k-1}} - v_{i_k}}{h_{i_{k-1}} - h_{i_k}}$ for all $k \in \{1, 2, \ldots, K\}$. From Definition 1 we know that customer classes $i_{k-1} + 1, i_{k-1} + 2, \ldots, i_k$ would join service grade (p_k, d_k) for all $k = 1, 2, \ldots, K$. Thus, choosing the K service grades is equivalent to choosing the indices $i_1, i_2, \ldots, i_K$, as per Definition 1. Moreover, the optimal menu would offer N incentive compatible service grades targeting each of the N customer classes. Thus, the worst case analysis formulation

for this problem is given by,

(1.11)

$$\overline{R}_N^K = \inf_{\lambda>0,v>0,h>0} \left\{ \frac{\max_{\{i_1,i_2,\ldots,i_K\}}\{\overline{\lambda}_{i_1} v_{i_1} + \sum_{k=2}^{K} (\overline{\lambda}_{i_k} - \overline{\lambda}_{i_{k-1}})\big(v_{i_k} - h_{i_k}\big(\frac{v_{i_{k-1}} - v_{i_k}}{h_{i_{k-1}} - h_{i_k}}\big)\big)\}}{\lambda_1 v_1 + \sum_{k=2}^{N} \lambda_k \big(v_k - h_k\big(\frac{v_{k-1} - v_k}{h_{k-1} - h_k}\big)\big)} \right\},$$

$$\text{s.t. } v_k > v_{k+1},\ \frac{h_k}{h_{k+1}} > \frac{v_k}{v_{k+1}},\ \text{for all } k = 1, 2, \ldots, N-1.$$

Again, since we are deriving an upper bound, we can simply use the solution for problem (1.9), which would be a feasible solution for (1.11). In particular, we let

$$\frac{\lambda_{i+1}}{\lambda_i} \to \infty, \frac{v_i}{v_{i+1}} \to \infty \text{ and } \frac{h_i}{h_{i+1}} \to \infty \text{ such that } \overline{\lambda}_i v_i = \kappa \text{ and } \frac{v_i / v_{i+1}}{h_i / h_{i+1}} \to 0$$

where κ is some constant. Using the conditions $\frac{h_i}{h_{i+1}} \to \infty$ and $\frac{v_i/v_{i+1}}{h_i/h_{i+1}} \to 0$ in (1.11) we get,

$$\overline{R}_N^K = \inf_{\lambda>0,v>0} \left\{ \frac{\max_{\{i_1,i_2,\ldots,i_K\}}\{\overline{\lambda}_{i_1} v_{i_1} + \sum_{k=2}^{K} (\overline{\lambda}_{i_k} - \overline{\lambda}_{i_{k-1}}) v_{i_k}\}}{\sum_{k=1}^{N} \lambda_k v_k} \right\}. \tag{1.12}$$

Additionally, condition $\frac{\lambda_{i+1}}{\lambda_i} \to \infty$ along with condition $\overline{\lambda}_i v_i = \kappa$ imply

$$\overline{R}_N^K = \inf_{\lambda>0,v>0} \left\{ \frac{\max_{\{i_1,i_2,\ldots,i_K\}}\{\sum_{k=1}^{K} \lambda_{i_k} v_{i_k}\}}{\sum_{k=1}^{N} \lambda_k v_k} \right\} = \frac{\max_{\{i_1,i_2,\ldots,i_K\}}\{\sum_{k=1}^{K} \kappa\}}{\sum_{k=1}^{N} \kappa} = \frac{K}{N}. \tag{1.13}$$

Theorem 1 implies that the WCRR converges to zero as the number of customer classes grows without bound. Note that we have, $\frac{R_N^{K+1}}{R_N^K} = 1 + \frac{1}{K}$ which implies that in the worst case scenario, marginal benefit of offering an additional service grade is the highest going from $K = 1$ to $K = 2$ and diminishes as the number of offered service grades, K, increases. Also, we note that the rate of convergence of R_N^K to zero is $\frac{1}{N}$, which implies that the firm loses a lot of revenue fairly quickly as the number of customer classes increase. This

happens since the worst case scenario for the firm is realized as the consecutive customer classes become infinitely more impatient with infinitely higher valuation, which minimizes the value of offering a simplified menu. Thus, the firm faces customer classes that are extremely heterogeneous in terms of their valuations, delay sensitivities and arrival rates.

Although customer classes could potentially differ from each other a lot in terms of their valuations, delay sensitivities and arrival rates, in practice, we expect that customer classes have limited heterogeneity. As a result, a natural question arises as to how the WCRR, would get affected if the valuations, delay sensitivities and the arrival rates are bounded. We consider this question in the subsequent sections and investigate the value of offering simplified menus with a limited number of service grades, under limited heterogeneity.

1.7. Worst Case Analysis Under Limited Heterogeneity: Delay Sensitivity

So far we have established that, the worst case scenario is realized when the firm faces customer classes that are infinitely more delay sensitive having infinitely higher valuations relative to one another. As a result, offering a simplified menu becomes less valuable with increasing number of customer classes and Theorem 1 in fact establishes that R_N^K converges zero as the number of customer classes grows without bound. Although customer classes could potentially differ from each other a lot in terms of their valuations, delay sensitivities and arrival rates, in practice, we expect that customer classes have limited heterogeneity. In this section, we focus on the worst case analysis when the arrival rates and the delay sensitivities for the customer classes are bounded (and hence the term *limited heterogeneity*, which we formalize in subsequent sections). First, in

§1.7.1 we analyze the effect of having bounded arrival rates, on the WCRR. We show that although the WCRR converges to zero, having bounded arrival rates slows down its rate of convergence. Following this, in §1.7.2 we present the worst case analysis for the firm's revenue when it offers a single service grade, i.e. $K = 1$, under limited heterogeneity in delay sensitivities and arrival rates. In particular, we assume that the ratio of the delay sensitivities of any two customer classes is bounded and arrival rates for the customer classes are fixed. Using the results in this section, we analyze the WCRR under bounded valuations and delay sensitivities in §1.8. In §1.7.2 and §1.8 we focus our worst case analysis on the single service grade case, i.e. $K = 1$. The reason for this, as stated earlier, is that it is complicated to characterize the structure of the worst case optimal menu for $K > 1$ (see Remark 1).

1.7.1. Bounded Arrival Rates

In this section, we analyze the effect of having bounded arrival rates on R_N^K. In particular, we use $m > 0$ to denote a lower bound on $\frac{\lambda_k}{\lambda_1}$, i.e. $m \leq \frac{\lambda_k}{\lambda_1}$ for all $k = 1, 2, \ldots, N$. We use $M < \infty$ to denote an upper bound on $\frac{\lambda_k}{\lambda_1}$, i.e. $\frac{\lambda_k}{\lambda_1} \leq M$ for all $k = 1, 2, \ldots, N$. If the arrival rates are bounded then the exact analysis of (1.9), the WCRR for $K = 1$, is possible. We evaluate R_N^1 by solving problem in (1.10) under the assumption of bounded arrival rates. Thus, we have,

$$
\begin{aligned}
R_N^1 = \inf_{\lambda_k > 0} & \left\{ \frac{1}{\sum_{k=1}^{N} \frac{\lambda_k}{\overline{\lambda}_k}} \right\} \\
& \text{s.t. } m \leq \frac{\lambda_k}{\lambda_1} \leq M \text{ for all } k = 1, 2, \ldots, N.
\end{aligned} \tag{1.14}
$$

where $\overline{\lambda}_k = \sum_{i=1}^{k} \lambda_i$ for all $k = 1, 2, \ldots, N$. The objective function in (1.14) can be rewritten as $\frac{1}{\sum_{k=1}^{N} \frac{\lambda_k/\lambda_1}{\overline{\lambda}_k/\lambda_1}}$. Applying the bounds on arrival rates, i.e. m and M, we have

$$R_N^1 = \frac{1}{1 + \sum_{k=2}^{N} \frac{M}{1+(k-1)m}}. \tag{1.15}$$

Next, for $K > 1$, we derive an upper bound to R_N^K using (1.11), and using the solution to (1.9) as a feasible solution to (1.11), similar to what we did in §1.6. In particular, we let

$$\frac{h_i}{h_{i+1}} \to \infty \text{ and } v_k = \frac{1}{\overline{\lambda}_k}.$$

Replacing these values in (1.11) we get the following upper bound:

$$\frac{\max_{\{i_1, i_2, \ldots, i_K\}} \{K - \sum_{k=2}^{K} \frac{\overline{\lambda}_{i_{k-1}}}{\overline{\lambda}_{i_k}}\}}{\sum_{k=1}^{N} \frac{\lambda_k}{\overline{\lambda}_k}} \leq \frac{K}{\sum_{k=1}^{N} \frac{\lambda_k}{\overline{\lambda}_k}}.$$

Using the fact that the arrival rates are bounded, we choose the values of $\frac{\lambda_k}{\overline{\lambda}_k}$ in the right hand side of the inequality such that it gives us the tightest bound. Thus we have,

$$R_N^K \leq \frac{K}{1 + \sum_{k=2}^{N} \frac{M}{1+(k-1)m}}. \tag{1.16}$$

The following proposition combines (1.15) and (1.16) to characterize the WCRR, R_N^K, under the assumption of bounded arrival rates.

Proposition 2. *If the arrival rates are bounded, i.e.* $m \leq \frac{\lambda_k}{\lambda_1} \leq M$ *for all* $k = 1, 2, \ldots, N$ *where* $m > 0$ *and* $M < \infty$*, the WCRR is such that,*

$$\frac{m/M}{\log(N) + \gamma_N} \leq R_N^K \leq \frac{K}{\log(N) + \gamma_N}$$

where $\gamma < \gamma_N < \gamma + \frac{1}{2}, \lim_{N\to\infty} \gamma_N \to \gamma$ *and* γ *is the Euler-Mascheroni constant.*

We note that the WCRR still converges to zero as the number of customer classes grows without bound. However, Proposition 2 says that if the arrival rates are bounded then the rate of convergence of R_N^K to zero slows down such that $R_N^K \propto \frac{1}{\log(N)+\gamma_N}$, as compared to $R_N^K \propto \frac{1}{N}$ when arrival rates are unbounded.

1.7.2. Bounded Delay Sensitivity and Fixed Arrival Rates

In the last section, we have observed that the worst case is realized as the ratio of consecutive delay sensitivities grows without bound. In this section, we perform the worst case analysis for the firm's revenue when it offers a single service grade under the assumption that the ratio of delay sensitivities of any two customer classes is bounded. In particular, since the consecutive classes become progressively more delay sensitive, we assume that $\frac{h_1}{h_N} \leq \Delta_h$ for some $\Delta_h > 1$, where $h_1 \geq h_2 \geq \cdots \geq h_N$. Furthermore, we assume that the arrival rates for all the customer classes are bounded and fixed. In particular, a fixed sequence $\{\lambda_k\}_{k=1}^N$ denotes the arrival rates for N customer classes, where $m \leq \frac{\lambda_k}{\lambda_1} \leq M$ for all $k = 1, 2, \ldots, N$ for some constant $m > 0$ and $M < \infty$. We maintain this assumption throughout the rest of this section unless otherwise specified. Thus, we are interested in

solving the following optimization problem:

$$R_N^1(\Delta_h) = \inf_{v>0,h>0} \left\{ \frac{\pi_N^1}{\pi_N^*} \right\} \quad \text{s.t. } \frac{h_1}{h_N} \leq \Delta_h,\ h_1 \geq h_2 \geq \cdots \geq h_N. \tag{1.17}$$

Lemma 1 establishes that the worst case optimal menu, WCOM, belongs to set $\mathbb{M}$ of menus which induces customer segmentation σ. WCOM, under unlimited heterogeneity offers full differentiation, i.e., all customer classes join service, and each service grade serves a single customer class. However, in the case with limited heterogeneity, this need not necessarily hold. To see this, we note that in order to provide a menu which differentiates all N customer classes, the necessary condition $\frac{h_1}{h_N} > \frac{\overline{\lambda}_N}{\overline{\lambda}_1}$, as given by Proposition 1(i) needs to hold. Therefore, if the value of Δ_h is such that $\Delta_h < \frac{\overline{\lambda}_N}{\overline{\lambda}_1}$, then we have, $\frac{h_1}{h_N} \leq \Delta_h < \frac{\overline{\lambda}_N}{\overline{\lambda}_1}$, which violates the condition in Proposition 1(i). This illustrates that the worst case optimal menu might not differentiate all N customer classes. As a result, depending on Δ_h, the optimal menu would result in pooling of some subset of customer classes in one service grade while differentiating some other customer classes.

In particular, for N customer classes, if the worst case optimal menu offers L service grades, then applying Definition 1 of customer segmentation $\sigma(i_1, i_2, \ldots, i_L)$, we can infer that indices $i_1, i_2, \ldots, i_L$ can be assigned in $\binom{N}{L} = \frac{N!}{L!(N-L)!}$ different ways. For e.g., if $N = 4$ and $L = 3$ then, depending on Δ_h, the worst case optimal menu could result in one of the following four customer segmentations: i) classes 1, 2 and 3 join separate service grades and class 4 does not join service, or all four classes join service but ii) classes 1 and 2 are pooled together, iii) classes 2 and 3 are pooled together, iv) classes 3 and 4 are pooled together into one service grade. Moreover, L could take a value in $\{2, 3, \ldots, N\}$ (L has to

be greater than 1, because the firm offers a single service grade, i.e. $K = 1$, and therefore the worst case would not be realized if the optimal menu offers a single service grade, i.e. $L = 1$). Thus, for N customer classes, $\sum_{L=2}^{N} \binom{N}{L} = 2^N - N - 1$ denotes the number of candidate customer segmentations that the worst case optimal menu could induce. We note that, the number of such candidate customer segmentations grow exponentially with N. However, Lemma 4 along with Definition 3 presents a characteristic of the worst case optimal menu which reduces the number of candidate customer segmentations.

Definition 3. *We define $\mathcal{M}_{ij}$ as a set of menus of service grades such that $\mathcal{M}_{ij} \subset \mathbb{M}_L$ (here $L = j - i + 1$), which results in customer segmentation $\sigma(i_1, i_2, \ldots, i_L)$ as per Definition 1, such that $i_1 = i, i_2 = i + 1, \ldots, i_L = j$ where, $1 \leq i < j \leq N$, and N denotes the number of customer classes.*

Lemma 4. *Under the optimal solution to (1.17), the optimal menu, WCOM, belongs to set $\mathcal{M}_{ij}$.*

Lemma 4 implies that the worst case is realized when the optimal menu segments all the customer classes into three subsets. The first subset comprises higher valuation classes $1, 2, \ldots, i$ that are pooled into a single service grade (classes are indexed in the decreasing order of their valuations). The second subset consists of classes $i + 1, i + 2, \ldots, j$ that are differentiated from each other and from all the pooled classes (i.e., each of the classes in this subset joins a service grade that is uniquely different from all other classes in terms of the offered price and the delays). The third subset is comprised of classes $j + 1, \ldots, N$ that are not served (i.e., these classes do not join any of the service grades). We note that i could take a value of 1, in which case there would be no pooled class. Similarly j could

take a value of N, in which case all the customer classes enter service. The number of ways in which i and j, in Definition 3, can be assigned is $\binom{N}{2} = \frac{N(N-1)}{2}$. Thus, Lemma 4 decreases the number of candidate customer segmentations under the worst case optimal menu from exponential to quadratic in N.

The revenue corresponding to any menu in set $\mathcal{M}_{ij}$ is denoted by $\pi(\mathcal{M}_{ij})$, which is given by

$$\pi(\mathcal{M}_{ij}) = \overline{\lambda}_i v_i + \lambda_{i+1} v_{i+1} \left(\frac{\delta_i - \frac{v_i}{v_{i+1}}}{\delta_i - 1} \right) + \cdots + \lambda_j v_j \left(\frac{\delta_{j-1} - \frac{v_{j-1}}{v_j}}{\delta_{j-1} - 1} \right),$$

where, $\delta_i = \frac{h_i}{h_{i+1}}$. In order to ensure that the offered prices are positive, we need the price condition, $\delta_k > \frac{v_k}{v_{k+1}}$ for all $k = i, i+1, \ldots, j-1$. Thus, i and j in addition to v and h are decision variables in the optimization problem (1.17), which can be rewritten as,

$$R_N^1(\Delta_h) = \min_{1 \le i < j \le N} r_N^{ij} \tag{1.18}$$

where

$$\begin{aligned} r_N^{ij} = \inf_{v_k > 0, \delta_k} & \left\{ \frac{\max_{k \in \{1,2,\ldots,N\}} \overline{\lambda}_k v_k}{\pi(\mathcal{M}_{ij})} \right\} \\ \text{s.t.} \ & \prod_{k=1}^{N-1} \delta_k \le \Delta_h, \\ & \delta_k \ge 1, \ v_k > v_{k+1} \text{ for all } k = 1, 2, \ldots, N-1, \\ & \delta_k > \frac{v_k}{v_{k+1}} \text{ for all } k = i, i+1, \ldots, j-1. \end{aligned} \tag{1.19}$$

In order to solve the optimization problem (1.18), first we focus on the solution of the inner optimization problem (1.19) which is equivalent to (1.18) holding i and j fixed. The following proposition characterizes the optimal valuations and delay sensitivities.

Proposition 3. *Under the optimal solution to (1.19),*

(i) Valuations, v_k, are such that, $1 \leq \frac{v_k}{v_{k+1}} \leq \frac{\overline{\lambda}_{k+1}}{\overline{\lambda}_k}$ for all $1 \leq k \leq i-1$, $\frac{v_k}{v_{k+1}} = \frac{\overline{\lambda}_{k+1}}{\overline{\lambda}_k}$ for all $i \leq k \leq j-1$ and $\frac{v_k}{v_{k+1}} \geq \frac{\overline{\lambda}_{k+1}}{\overline{\lambda}_k}$ for all $j \leq k \leq N-1$.

(ii) Delay sensitivities, h_k, are such that they satisfy the equation: $\prod_{k=i}^{j-1} \delta_k = \Delta_h$ where $\delta_k = \frac{h_k}{h_{k+1}}$. Moreover, $\delta_k = 1$ for all $1 \leq k \leq i-1$ and $j \leq k \leq N-1$, and $\delta_i, \delta_{i+1}, \ldots, \delta_{j-1}$ are such that they satisfy the set of implicit equations: $\frac{\delta_i}{c_i(c_i+1)(\delta_i-1)^2} = \frac{\delta_{i+1}}{c_{i+1}(c_{i+1}+1)(\delta_{i+1}-1)^2} = \cdots = \frac{\delta_{j-1}}{c_{j-1}(c_{j-1}+1)(\delta_{j-1}-1)^2}$, where, $c_k = \frac{\overline{\lambda}_k}{\lambda_{k+1}}$ and $\overline{\lambda}_k = \sum_{i=1}^{k} \lambda_i$.

For $N = 4$, by using Lemma 4, we see that there are $\frac{N(N-1)}{2} = 6$ possible customer segmentations under the worst case optimal menu. Figure 1.1 depicts the WCRR for each of these possible customer segmentations that could be induced by the optimal menu, which are obtained by solving (1.17) multiple times, each time holding the customer segmentation fixed.

The benefit of differentiating any two customer classes comes from the fact that the higher valuation customer class can be charged a higher price as compared to the lower valuation class. For the firm to be able to differentiate two customer classes, the service grade joined by the lower valuation class should offer higher delay as compared to the service grade joined by the higher valuation customer class. As these two customer

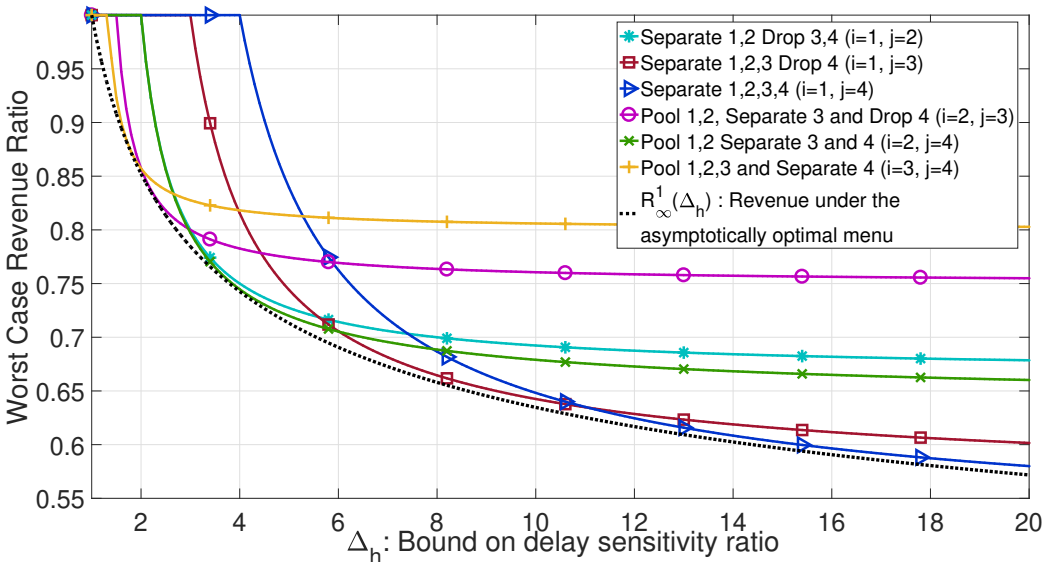


Figure 1.1. WCRR corresponding to all possible customer segmentations induced by the optimal menu, with number of customer classes, $N = 4$, under bounded delay sensitivities, i.e. $\frac{h_1}{h_4} \leq \Delta_h$. For any given Δ_h, the customer segmentation that results in the least value of the WCRR is the customer segmentation that is induced by the optimal menu under the solution to problem (1.17). For illustration purposes we set the arrival rates for all classes to one, i.e., $\lambda_k = 1$ for all $k = 1, 2, \ldots, N$.

classes become less delay sensitive relative to one another, the offered delay that is required to differentiate these customer classes increases and vice versa. The expression $\prod_{k=i}^{j-1} \delta_k = \Delta_h$ from Proposition 3(ii) implies that, as higher number of customer classes are differentiated, i.e., as $j - i$ gets larger, δ_k gets smaller and consequently the offered delay becomes larger. As the offered delay become larger, the price charged for the service grade become smaller. This highlights a trade-off involved in problem (1.17) between offering delay-differentiation to a higher number of customer classes and pooling multiple customer classes into a single service grade. This trade-off is not relevant when delay sensitivities are unbounded. We recall that the optimal solution to (1.8) is achieved as

the ratio of consecutive delay sensitivities grow without bound. For bounded delay sensitivities this trade-off becomes relevant, and indeed we observe this in Figure 1.1. We note that for higher values of Δ_h, the customer segmentation that results in the least value of WCRR, is such that it differentiates a higher number of customer classes. On the other hand, if Δ_h is lower, customer segmentations that pool multiple customer classes together, result in lower values of the WCRR.

Now, we consider the objective function in the optimization problem (1.19), which is given by,

$$\frac{\max_{k\in\{1,2,\ldots,N\}} \overline{\lambda}_k v_k}{\overline{\lambda}_i v_i + \lambda_{i+1} v_{i+1}\left(\frac{\delta_i - \frac{v_i}{v_{i+1}}}{\delta_i - 1}\right) + \cdots + \lambda_j v_j \left(\frac{\delta_{j-1} - \frac{v_{j-1}}{v_j}}{\delta_{j-1} - 1}\right)}. \tag{1.20}$$

The valuations under the optimal solution to (1.19), as given in Proposition 3(i), imply that the maximum value in the numerator of the objective function in (1.20) is achieved for indices $i \le k \le j-1$. By choosing $k = i$ and dividing both numerator and denominator by $\overline{\lambda}_i v_i$, we can rewrite (1.20) as,

$$\frac{1}{1 + \sum_{k=i}^{j-1} \frac{\lambda_{k+1} v_{k+1}}{\overline{\lambda}_i v_i}\left(1 + \frac{1 - \frac{v_k}{v_{k+1}}}{\delta_k - 1}\right)}.$$

Defining $c_k := \frac{\overline{\lambda}_k}{\lambda_{k+1}}$ and applying Proposition 3(i) we get, $\frac{v_k}{v_{k+1}} = \frac{\overline{\lambda}_{k+1}}{\overline{\lambda}_k} = 1 + \frac{1}{c_k}$ for all $k = i, i+1, \ldots, j-1$. By expressing $\frac{v_k}{v_{k+1}}$ in terms of c_k and rearranging the terms we can rewrite (1.20), which is given by,

$$\frac{1}{1 + \sum_{k=i}^{j-1}\left(\frac{1}{1+c_k} - \frac{1}{c_k(c_k+1)(\delta_k - 1)}\right)}.$$

Thus, using the results from Proposition 3, we have expressed the objective function in r_N^{ij}, as given in (1.19), in terms of just δ_k as opposed to v_k and h_k. Since, we consider arrival rates $\{\lambda_k\}_{k=1}^N$ as fixed, c_k defined as $c_k := \frac{\overline{\lambda}_k}{\lambda_{k+1}}$ are constants in the optimization problem (1.18). Solving (1.18) for any finite N is analytically intractable, hence, we consider optimization problem (1.18) as the number of customer classes, N, grows without bound. Thus, we are interested in solving the following optimization problem:

$$R_\infty^1(\Delta_h) = \min_{1\le i<j} \inf_{\delta_k} \left\{ \frac{1}{1+\sum_{k=i}^{j-1}\left(\frac{1}{1+c_k} - \frac{1}{c_k(c_k+1)(\delta_k-1)}\right)} \right\} \tag{1.21}$$
$$\text{s.t} \prod_{k=i}^{j-1} \delta_k \le \Delta_h,\ \delta_k > 1 + \frac{1}{c_k} \text{ for all } k = i, i+1, \ldots, j-1.$$

We note that the constraint $\delta_k > 1 + \frac{1}{c_k}$ for all $k = i, i+1, \ldots, j-1$ is the price condition. Reformulation (1.21) in terms of c_k will be particularly useful going forward in the next section when we consider bounded valuations. The following Theorem provides the solution to (1.21) and thus, characterizes the WCRR and the asymptotically optimal menu.

Theorem 2. *(i) For any given $\Delta_h > 1$, such that $\frac{h_1}{h_N} \le \Delta_h$ where $h_1 \ge h_2 \ge \cdots \ge h_N$, and for any given bounded sequence of arrival rates, the worst case revenue ratio (WCRR) for a firm offering a single service grade, $K = 1$, as the number customer classes grows without bound, i.e. $N \to \infty$, is given by,*

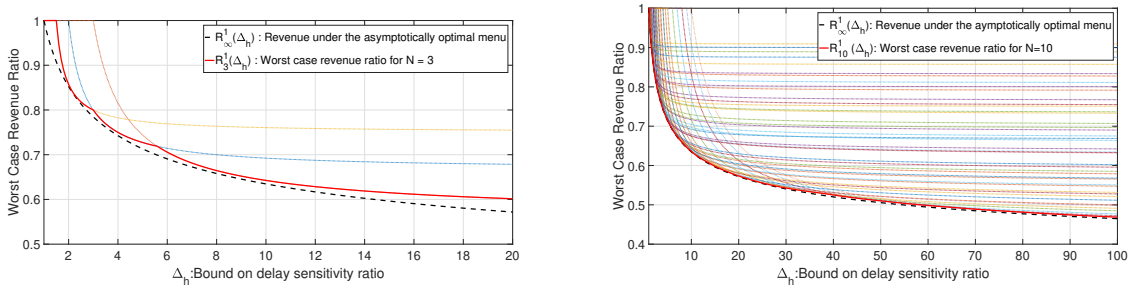
$$R_\infty^1(\Delta_h) = \frac{1}{1 + \frac{1}{4}\log(\Delta_h)}.$$

(ii) $R_N^1(\Delta_h)$ is bounded from below by $R_\infty^1(\Delta_h)$ for all finite N.

(iii) The worst case optimal menu, WCOM, offers infinitely many differentiated service grades, i.e. $(j^* - i^*) \to \infty$. *For the special case where all the arrival rates are equal, i.e. if* $\lambda_k = 1$ *for all* $k = 1, 2, \ldots, \infty$, *the optimal indices* i^* *and* j^* *satisfy* $i^* = \dfrac{j^*}{\sqrt{\Delta_h}}$, *and the worst case delay sensitivity ratios are* $\delta_k^* = 1 + \dfrac{2}{k+1}$ *for all* $i^* \leq k \leq j^* - 1$.

As the number of customer classes increase, by offering a simplified menu, the firm can only get worse off in terms off its revenue loss, i.e. the WCRR can only (weakly) decrease. Hence for any given Δ_h, the firm's WCRR under the asymptotically optimal menu, $R^1_\infty(\Delta_h)$, provides a lower bound to $R^1_N(\Delta_h)$ (see Figure 1.2).

Optimization problem (1.21) assumes that there is a fixed sequence, $\{\lambda_k\}_{k=1}^{\infty}$, of arrival rates which implies that $c_k = \frac{\lambda_k}{\lambda_{k+1}}$ are constants in (1.21). Moreover, since we have assumed that the arrival rates are bounded, i.e. $m \leq \frac{\lambda_k}{\lambda_1} \leq M$, this implies that $\frac{m}{M}k \leq$



(a) Number of classes, N = 3

(b) Number of classes, N = 10

Figure 1.2. Gap between the WCRR for finite number of customer classes, N, and the asymptotically optimal menu under equal arrival rates. The dotted lines represent the worst case revenue ratios corresponding to all candidate customer segmentations induced by the optimal menu. We do not enumerate the dotted lines as the number of dotted lines increase quadratically with N.

$c_k \leq \frac{M}{m}k$ for all $k = 1, 2, \ldots, \infty$. To see this, we recall that c_k is defined as $c_k = \frac{\overline{\lambda}_k}{\lambda_{k+1}} = \frac{\sum_{n=1}^{k} \lambda_n/\lambda_1}{\lambda_{k+1}/\lambda_1}$. Thus, using the bounds on the arrival rates, we have $\frac{m}{M}k \leq \frac{\sum_{n=1}^{k} \lambda_n/\lambda_1}{\lambda_{k+1}/\lambda_1} \leq \frac{M}{m}k$. Hence, $\frac{m}{M}k \leq c_k \leq \frac{M}{m}k$ for all $k = 1, 2, \ldots, \infty$. However, it is important to note that $R^1_\infty(\Delta_h)$, as given by Theorem 2, does not depend either on m or on M. Thus, the bounds on the arrival rates do not affect the WCRR. Furthermore, we also note that the arrival rates, λ_k, do not appear in the expression for $R^1_\infty(\Delta_h)$ for any $k = 1, 2, \ldots, \infty$. This observation leads to the following Corollary.

Corollary 1. *For any given bounded sequence of arrival rates $\{\lambda_k\}_{k=1}^{\infty}$, the optimal solution to problem (1.21) does not depend on λ_k, and hence it does not depend on c_k, for any $k \in \{1, 2, \ldots, \infty\}$.*

Proposition 3(i) links the arrival rates and the valuations in the worst case analysis such that $\frac{v_k}{v_{k+1}} = \frac{\overline{\lambda}_{k+1}}{\overline{\lambda}_k}$ for $i \leq k \leq j-1$, where $\frac{\overline{\lambda}_{k+1}}{\overline{\lambda}_k}$ is expressed in terms of c_k as $\frac{\overline{\lambda}_{k+1}}{\overline{\lambda}_k} = \left(1 + \frac{1}{c_k}\right)$. Using these two relations we arrive at the following equality,

$$\frac{v_i}{v_j} = \prod_{k=i}^{j-1} \frac{v_k}{v_{k+1}} = \prod_{k=i}^{j-1} \left(1 + \frac{1}{c_k}\right). \tag{1.22}$$

The optimal solution to (1.21) is achieved as both decision variables i and j grow without bound such that $\prod_{k=i}^{j-1} \left(1 + \frac{1}{c_k}\right)$ approaches $\sqrt{\Delta_h}$ (see proof of Theorem 2). For the special case of equal arrival rates, i.e. if $\lambda_k = 1$ for all $k = 1, 2, \ldots, \infty$, we get $c_k = \frac{\overline{\lambda}_k}{\lambda_{k+1}} = k$, which implies that $\prod_{k=i}^{j-1} \left(1 + \frac{1}{c_k}\right) = \prod_{k=i}^{j-1} \frac{k+1}{k} = \frac{j}{i}$. This establishes the optimal values i^* and j^* satisfy $\frac{j^*}{i^*} = \sqrt{\Delta_h}$, as presented in Theorem 2(iii). (For a quick check, we note from Proposition 3(ii) that, $\prod_{k=i}^{j-1} \delta_k = \frac{h_i}{h_j} = \Delta_h$. Moreover, from Proposition 1 we have $\frac{h_i}{h_j} > \frac{\overline{\lambda}_j}{\overline{\lambda}_i} = \frac{j}{i}$ since classes i and j are differentiated. Combining these we have,

$\Delta_h = \frac{h_i}{h_j} > \frac{j}{i} = \sqrt{\Delta_h}$.) In general, under the optimal solution to (1.21), the specific relationship between i^* and j^* would depend on $\{c_k\}_{k=1}^{\infty}$, and thus it would depend on the sequence of arrival rates, $\{\lambda_k\}_{k=1}^{\infty}$.

Now, Corollary 1 implies that if we relax the assumption of fixed arrival rates and make c_k a decision variable in (1.21), for all $k = i, i+1, \ldots, j-1$, we would get the following equivalent formulation:

$$\begin{aligned} &\min_{1\le i<j} \inf_{\delta_k, c_k} \left\{ \frac{1}{1+\sum_{k=i}^{j-1}\left(\frac{1}{1+c_k} - \frac{1}{c_k(c_k+1)(\delta_k-1)}\right)} \right\} \\ &\text{s.t } \prod_{k=i}^{j-1} \delta_k \le \Delta_h,\ \delta_k > 1 + \frac{1}{c_k},\ c_k > 0 \text{ for all } k = i, i+1, \ldots, j-1. \end{aligned} \tag{1.23}$$

We note that in problem (1.23), we can rename the decision variables c_k and δ_k as c_{k-i+1} and δ_{k-i+1} respectively, for all $k = i, i+1, \ldots, j-1$, without altering the optimization problem. Thus, we rename decision variables $\{c_i, c_{i+1}, \ldots, c_{j-1}\}$ to $\{c_1, c_2, \ldots, c_{j-i}\}$. Similarly, we rename decision variables $\{\delta_i, \delta_{i+1}, \ldots, \delta_{j-1}\}$ to $\{\delta_1, \delta_2, \ldots, \delta_{j-i}\}$. In addition, we have established in Theorem 2(iii) that the the WCOM offers infinitely many differentiated service grades, i.e. $(j^* - i^*) \to \infty$. Hence, we can rewrite (1.23) as follows:

$$\begin{aligned} &\inf_{\delta_k, c_k} \left\{ \frac{1}{1+\sum_{k=1}^{\infty}\left(\frac{1}{1+c_k} - \frac{1}{c_k(c_k+1)(\delta_k-1)}\right)} \right\} \\ &\text{s.t } \prod_{k=1}^{\infty} \delta_k \le \Delta_h,\ \delta_k > 1 + \frac{1}{c_k},\ c_k > 0 \text{ for all } k = 1, 2, \ldots, \infty. \end{aligned} \tag{1.24}$$

As already mentioned, the optimal value in problem (1.21) is achieved as $\prod_{k=i^*}^{j^*-1}\left(1+\frac{1}{c_k}\right)$ approaches $\sqrt{\Delta_h}$. Since formulation (1.24) is equivalent to (1.21), it implies that the

optimal value of (1.24) is achieved as $\prod_{k=1}^{\infty}\left(1+\frac{1}{c_k}\right)$ approaches $\sqrt{\Delta_h}$. Equivalently, $\frac{v_1}{v_\infty}$ approaches $\sqrt{\Delta_h}$, as established by the equality in (1.22). Finally, this leads us to investigate the effect of having bounded valuations on the WCRR, which is what we consider in the next section.

1.8. Worst Case Analysis Under Limited Heterogeneity: Delay Sensitivity and Valuation

In this section, we analyze the effect of bounded valuations and bounded delay sensitivities on the WCRR when the firm offers a single service grade, i.e. $K = 1$. First, in §1.8.1 we provide the asymptotic lower bound to the WCRR as the number of customer classes, N, grows without bound. Following this, in §1.8.2, we numerically compute the WCRR for finitely many customer classes and compute the gap with respect to the asymptotic lower bound.

1.8.1. Asymptotic Lower Bound

We are interested in solving problem (1.24) with the added constraint $\frac{v_1}{v_\infty} \leq \Delta_v$ such that $v_k > v_{k+1}$ for all $k = 1, 2, \ldots, \infty$, where $\Delta_v > 1$ represents the bound on the valuations. As previously mentioned, the constraint $\frac{v_1}{v_\infty} \leq \Delta_v$ is equivalent to $\prod_{k=1}^{\infty}\left(1+\frac{1}{c_k}\right) \leq \Delta_v$ as established by the equality in (1.22). Thus, we are interested in solving the following

problem:

$$R^1_\infty(\Delta_h,\Delta_v)=\inf_{\delta_k,c_k}\left\{\frac{1}{1+\sum_{k=1}^{\infty}\left(\frac{1}{1+c_k}-\frac{1}{c_k(c_k+1)(\delta_k-1)}\right)}\right\}$$

$$\text{s.t. } \prod_{k=1}^{\infty}\delta_k\le\Delta_h, \tag{1.25}$$

$$\prod_{k=1}^{\infty}\left(1+\frac{1}{c_k}\right)\le\Delta_v,$$

$$\delta_k>1+\frac{1}{c_k},\ c_k>0 \text{ for all } k=1,2,\ldots,\infty.$$

We note that since c_k is a decision variable in (1.25) for all $k=1,2,\ldots,\infty$, we are essentially optimizing over arrival rates. Thus, for solving (1.25) we do not impose the assumption of bounded arrival rates, that we maintained in section §1.7.2. The Theorem below provides the solution to optimization problem (1.25), which is the asymptotic lower bound for the WCRR when the firm provides a single service grade under limited heterogeneity in valuations and delay sensitivities.

Theorem 3. (Asymptotic Lower Bound) *For any given $\Delta_h>1$ and $\Delta_v>1$ such that $\frac{h_1}{h_N}\le\Delta_h$ and $\frac{v_1}{v_N}\le\Delta_v$, where $h_1>h_2>\cdots>h_N$ and $v_1>v_2>\cdots>v_N$, and the number of customer classes $N\to\infty$, the WCRR for a firm offering a single service grade, i.e., $K=1$ is:*

$$R^1_\infty(\Delta_h,\Delta_v)=\begin{cases}\dfrac{1}{1+\dfrac{\log(\Delta_v)}{\log(\Delta_h)}\log\left(\dfrac{\Delta_h}{\Delta_v}\right)} & \Delta_v<\sqrt{\Delta_h},\\[2ex] \dfrac{1}{1+\dfrac{1}{4}\log(\Delta_h)} & \Delta_v\ge\sqrt{\Delta_h}.\end{cases}$$

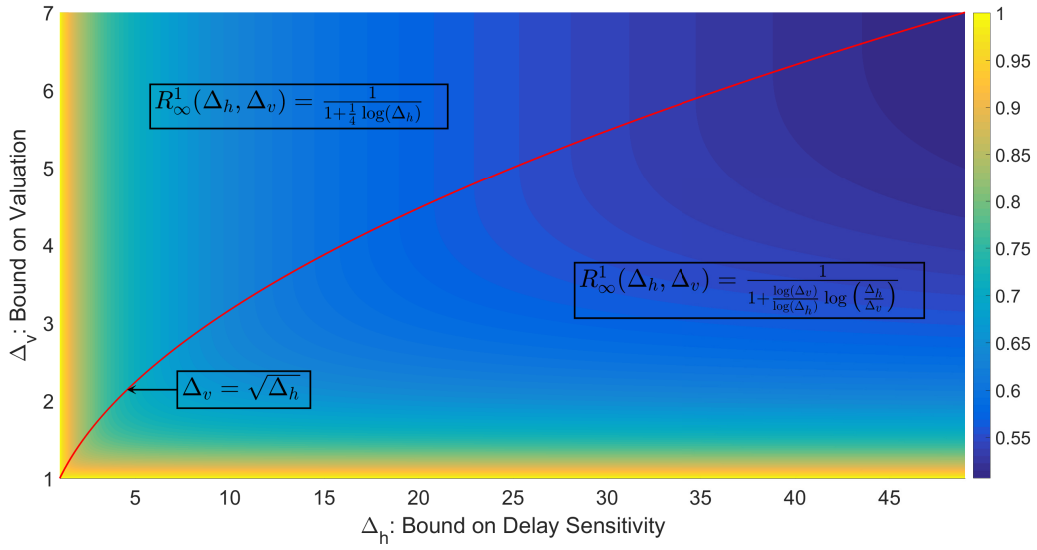


Figure 1.3. The asymptotic lower bound for WCRR, $R^1_\infty(\Delta_h, \Delta_v)$, as a function of the delay sensitivity bound, Δ_h, and the valuation bound, Δ_v. The curve $\Delta_v = \sqrt{\Delta_h}$ separates the two regions with different functional forms. The color scale represents the value of $R^1_\infty(\Delta_h, \Delta_v)$.

Theorem 3 present the asymptotic lower bound to the WCRR for a firm offering a single service grade under the assumption of limited heterogeneity. We note that $R^1_\infty(\Delta_h, \Delta_v)$ is bounded from below away from zero. This suggests that offering a single service grade, as opposed to a menu of multiple service grades, can be valuable and the firm would be guaranteed a fraction of the optimal revenue. We study the dependence of Δ_v and Δ_h on $R^1_\infty(\Delta_h, \Delta_v)$ in the numerical study (see §1.8.2) and show that $R^1_\infty(\Delta_h, \Delta_v)$ changes very slowly. For any given Δ_v such that $\Delta_v < \sqrt{\Delta_h}$, we see that $R^1_\infty(\Delta_h, \Delta_v) \to \frac{1}{1+\log(\Delta_v)}$ as $\Delta_h \to \infty$ which is the WCRR under bounded valuations and unbounded delay sensitivities. Similarly, for any given Δ_h, by letting $\Delta_v \to \infty$, we can recover the result in Theorem 2(i), the WCRR under bounded delay sensitivities.

Another interesting observation from Theorem 3 and Figure 1.3 is that for any $\Delta_v \geq \sqrt{\Delta_h}$, changes in Δ_v have no effect on the WCRR. This happens as a result of the price condition. For the firm to be able to differentiate customer classes by providing service grades k and $k+1$ with non-negative prices, it is necessary that $\frac{v_k}{v_{k+1}} < \frac{h_k}{h_{k+1}}$. This is the reason, why the bound on valuations, Δ_v, does not affect our worst case analysis when $\Delta_v \geq \sqrt{\Delta_h}$.

Remark 1. *(*Offering Multiple Service Grades*) In this paper, our goal has been to analyze the worst case revenue loss by a firm offering a simplified menu of service grades. This analysis relies on the structure of the worst case optimal menu when the firm offers a single service grade, i.e. $K = 1$. We have established this structure in Lemmas 1 and 4. However, trying to establish the structure for the worst case optimal menu when the firm offers multiple service grades, i.e. $K > 1$, becomes complicated. We can use the result in Theorem 3 to establish a lower bound to $R^K_\infty(\Delta_h, \Delta_v)$. In particular, we have*

$$R^1_\infty(\Delta_h, \Delta_v) \leq R^K_\infty(\Delta_h, \Delta_v).$$

A firm offering a menu with more than one service grade can always choose to offer a single service grade, if it were optimal to do so. Hence, by having the option of offering more than one service grade, the firm cannot be worse-off in terms of maximizing its revenue. Therefore $R^K_\infty(\Delta_h, \Delta_v)$ can only (weakly) increase as K increases.

1.8.2. Computational Results

For any finite N, as Δ_h and Δ_v grow without bound, $R^1_N(\Delta_h, \Delta_v)$ approaches $R^1_N = \frac{1}{N}$, as established in Theorem 1. Consequently, $R^1_\infty(\Delta_h, \Delta_v)$ should approach zero as Δ_h and Δ_v grow without bound. In this section, we numerically compute $R^1_N(\Delta_h, \Delta_v)$, for a set of values of Δ_h, Δ_v and N, and illustrate the convergence of $R^1_N(\Delta_h, \Delta_v)$ to its asymptotic lower bound $R^1_\infty(\Delta_h, \Delta_v)$, as presented in Theorem 3. In particular we compute the following gap,

$$R^1_N(\Delta_h, \Delta_v) - R^1_\infty(\Delta_h, \Delta_v).$$

From Theorem 3, we note that $R^1_\infty(\Delta_h, \Delta_v)$ does not change for $\Delta_v \geq \sqrt{\Delta_h}$. With this in mind, we pick the values of Δ_v such that it is comparable to $\sqrt{\Delta_h}$. From Table 1.8.1 we observe that, for a given value of N, the gap between $R^1_N(\Delta_h, \Delta_v)$ and $R^1_\infty(\Delta_h, \Delta_v)$

Δ_h	Δ_v	$R^1_3(\Delta_h, \Delta_v)(\%)$	$R^1_5(\Delta_h, \Delta_v)(\%)$	$R^1_{10}(\Delta_h, \Delta_v)(\%)$	$R^1_\infty(\Delta_h, \Delta_v)(\%)$
5	2	71.97	71.77	71.72	71.70
5	3	71.58	71.38	71.33	71.31
10	2	67.86	67.49	67.39	67.36
10	4	64.09	63.63	63.50	63.47
20	3	60.00	59.23	59.02	58.97
20	5	58.28	57.46	57.23	57.18
60	5	52.55	51.10	50.69	50.59
60	8	51.48	49.95	49.52	49.42
100	12	49.04	47.16	46.62	46.48
200	15	46.30	43.89	43.19	43.02
500	25	43.44	40.32	39.39	39.16

Table 1.1. Gap between the WCRR for finite number of customer classes under limited heterogeneity, $R^1_N(\Delta_h, \Delta_v)$ for $N = 3, 5$ & 10, and the asymptotic lower bound, $R^1_\infty(\Delta_h, \Delta_v)$.

increases as Δ_h and Δ_v increase. However, this gap is fairly small as compared to $\frac{1}{N}$, and grows fairly slowly with Δ_h and Δ_v. Moreover, for given values of Δ_h and Δ_v, the gap closes significantly faster as the number of customer classes, N, increase. Thus, the value of N need not be too large for the gap between $R^1_N(\Delta_h, \Delta_v)$ and $R^1_\infty(\Delta_h, \Delta_v)$ to be significantly small even for fairly large values of Δ_h and Δ_v.

1.9. Conclusion

We consider a revenue-maximizing firm which caters to price and time-sensitive customers by offering a menu of service grades. In order to limit the complexity of the offered menu, in practice, firms offer a menu with fewer service grades than the number of customer classes. Motivated by this, we quantify the firm's worst case revenue loss, incurred by offering a simplified menu of service grades as compared to the revenue-maximizing optimal menu. We characterize this gap in revenue by evaluating the minimum value of the ratio of the firm's revenue under a simplified menu and the revenue under the optimal menu, the WCRR. We show that the WCRR converges to zero as the number of customer classes grow without bound which indicates significant loss of revenue by the firm. Moreover, this worst case is realized when there is unlimited heterogeneity among the customer classes. In particular, consecutive customer classes are infinitely more delay sensitive with infinitely higher valuation for service, which minimizes the value of offering a simplified menu. Thus, as the number of customer classes grow without bound, any menu offering a finite number of service grades would be insufficient in capturing any given fraction of the optimal revenue.

Although customer classes could potentially differ from each other a lot in terms of their valuations, delay sensitivities and arrival rates, in practice, we expect that customer classes have limited heterogeneity. Hence, we analyze the WCRR under limited heterogeneity. In the unlimited heterogeneity setting, the worst case optimal menu results in a customer segmentation such that all customer classes are differentiated from each other. In contrast, in the limited heterogeneity setting the customer segmentation that is induced by the worst case optimal menu depends on the limited heterogeneity bounds on the delay sensitivity and valuation. The number of such candidate customer segmentations grows exponentially with the number of customer classes. This poses a challenge in analyzing the limited heterogeneity setting. We overcome this challenge by characterizing the worst case optimal menu and showing that it induces a customer segmentation which pools a subset of customer classes with higher valuations, while differentiating a subset of customers classes with lower valuations. This characterization of the worst case optimal menu allows us to reduce the growth of the number of candidate customer segmentations from exponential to quadratic in the number of customer classes. We show that, under limited heterogeneity, even if the number of customer classes grow without bound, the WCRR for the firm has a lower bound that is strictly greater than zero. This suggests that offering simplified menus could be valuable, and in doing so, the firm could recover a significant portion of the optimal revenue. Using Taylor series arguments, we provide an asymptotic lower bound to the WCRR as the number of customer classes grow without bound. The asymptotic lower bound is fairly robust to changes in the bounds on valuations and delay sensitivities. Characterization of the firm's worst case revenue loss in

terms of a measure of heterogeneity can be used to guide decision making when offering a simplified menu of service grades.

CHAPTER 2

The Queue Behind the Curtain: Information Disclosure in Omnichannel Services (Joint work with Achal Bassamboo and Martin Lariviere)

2.1. Introduction

How do you order a coffee? For many tech-savvy customers, the answer is to pull out their phone and use an app. Quick service restaurants (QSRs) such as Starbucks, Dunkin' and Chipotle have developed innovative online technology that allows customers to order whenever and from wherever they want. Online channels offer a convenient and streamlined ordering experience that aims to reduce the time customers wait in the store by letting them order before they arrive. By maintaining two channels, firms create a choice for tech-savvy customers. They may use the *online channel* to order before reaching the store or they may use the conventional *offline channel* and order after traveling to the physical store. With increased smartphone penetration, increasing number of customers are shifting from offline to online. In the third fiscal quarter of 2019, Chipotle's digital sales rose 88%, which constitutes about 18% of the total sales for the chain, with some stores pushing 30% (see Kelso, 2019).

Due to the novelty of these *omnichannel* systems, there are many open questions about their design and their implications for firms and the customers. For example, how do customers decide whether to order online or offline? Customers, as utility maximizers,

are presumably strategic in their channel choice. They are time-sensitive and prefer to minimize their time at the store. While delay sensitivity is an important factor, in the context of QSRs, there is another significant consideration that goes into customers' choice of channel: quality. Imagine a customer ordering a latte via the Starbucks app. Assuming she prefers a hot drink, she would not want her latte prepared long before she arrives to pick it up. Similarly, a Chipotle customer would prefer not to have a cold burrito. These examples highlight the fact that customers, while being time-sensitive, are also sensitive to potential degradation in product quality. This results in a trade-off when deciding whether or not to order using the mobile app; less time spent waiting may result in a degraded product. We explore this trade-off in our paper.

Another open issue relates to the design of omnichannel systems: Whether to disclose congestion information to online customers? Some firms (e.g., Starbucks) provide app users with information such as expected wait-time, which signals the current state of the system. If queue information is provided through the app, tech-savvy customers have the possibility to sample the queue before arriving at the physical store. They could either order online immediately or they could wait until they arrive at the store and then order if they see a favorable evolution of the queue. Customers' choice of channel depends on their knowledge of the level of congestion in the system. System congestion, in turn, depends on the cumulative behavior resulting from individual customer decisions. In the absence of such real-time information, self-interested customers' decision is based on their rational expectations about the steady state behavior of the system. Thus, for an accurate evaluation of the performance of these omnichannel systems, it is essential to address the issue of information disclosure.

In this paper we address these questions regarding the design of omnichannel services. Specifically, we study the following:

(1) How does customer ordering strategy change as the service system moves from a single channel to an omnichannel system?
(2) Should service providers disclose current congestion levels (e.g. the queue length) to customers before they decide whether to place an online order?
(3) Are firm's better-off, in terms of revenue, by running multiple channels?
(4) Is an online option better for customers, in terms of increasing consumer surplus?

We consider two types of customers, app users and the non-app users. App users are tech-savvy customers who could either order remotely (online) or walk into the physical store to order (offline). In contrast, non-app users never use the online option; they visit the store and decide whether or not to enter the queue. The non-app user group could be comprised of, but not limited to, customers who are not tech-savvy or customers who do not frequent the store often enough to have the mobile app installed on their phones. If online ordering option is not offered, then app users, like non-app users, choose the offline option.

We adopt a game-theoretic discrete-time queueing framework to model the strategic behavior of customers in omnichannel systems, both with and without information disclosure to online customers. A single virtual queue holds orders from both customer types. All the orders (from app users and non-app users) are fulfilled in a first-come-first-served manner. Additionally, consistent with the literature (see Baron et al., 2019), we assume that the entire order queue is visible to all store customers (both app users, when they

arrive at the store, and non-app users). This assumption is motivated by firms, like McDonald's, using digital boards to display the number of orders in the system that are pending and being currently worked on.

For non-app users, we assume that they arrive to the market by physically appearing at the store. In contrast an app user arrives to the market when she is physically away from the store, but she will reach the store in the following time period. This customer faces a choice. She could place an order online; in which case her order may be prepared in parallel with her travels. Alternatively, she could wait until she reaches the store then decide whether to order or not to order. Each class of customer has a type-dependent wait-sensitivity. Additionally, app users are also quality sensitive. App users must then balance waiting and quality costs. If the firm discloses queue information, we find that the app users follow a queue-length dependent *dual-threshold policy* where they alternate between walking in (to avoid the quality penalty), ordering online (to shorten their wait), and walking in (to avoid a really long wait and potentially balking) as the queue grows. In the absence of online queue length information, app users' ordering decision is based on their steady-state expectation of the system. In this case, there are three possible customer behavior under equilibrium: all app users either order online, choose the offline option or potentially randomize between these two actions.

We find that the relative wait-sensitivities of app users and non-app users play a significant role in determining which channel arrangement delivers highest overall system throughput. Non-app users get crowded out by app users in omnichannel systems. If non-app users are relatively more wait-sensitive, so many of them may balk that the overall throughput in an omnichannel system is less than the throughput of a single

channel system. On the other hand, if app users are more wait-sensitive, then omnichannel systems outperform the single channel system. Moreover, disclosing queue information in an omnichannel system results in higher throughput only if app users are either highly quality sensitive or highly wait-sensitive, or if the system is heavily congested. Offering more visibility to app users lures them to place order in unfavorable states of the queue, in which they otherwise would not have joined the system if information were withheld. Thus, from the firm's perspective, we find that no channel arrangement dominates the other two in terms of throughput.

From the perspective of customers, we find that non-app users are consistently better-off in a single channel system; when online ordering is not offered, non-app users do not get crowded out and hence their consumer surplus is higher. Interestingly, we find that the app users might not necessarily benefit from an online ordering option, specifically if congestion information is withheld. This result illustrates the trade-off between visibility of system congestion, which is offered by the single channel, and the benefit of ordering ahead in the omnichannel system. We show that this trade-off is resolved and consumer surplus for omnichannel system increases if the firm discloses queue information to app users. We find that, depending on the relative proportion of non-app users and app users in the system, either single channel or omnichannel system could deliver higher overall combined consumer surplus. It is certainly possible that both segments are worse-off when online ordering is offered.

The key take-away from our paper is that the non-app users are worse-off when online ordering option is offered. Thus, the overall performance of an omnichannel system, both in terms of firm's revenue and the quality of service experienced by customers, highly

depends on the customer primitives and the relative proportion of non-app users and app users in the system.

2.1.1. Related Literature

Our paper contributes to the literature on the role of information in queues and the rapidly evolving literature on omnichannel services. The seminal paper by Naor (1969) studies the throughput and social welfare implications of self-interested customer behavior in an observable queue. The importance of what information is available to customers in queuing systems was highlighted by Edelson and Hilderbrand (1975), who studied the classic problem of Naor (1969) in an unobservable queue. Observable and unobservable systems are compared in terms of social welfare in Hassin (1986), and in terms of throughput in Chen and Frank (2004). There has been extensive work on this topic, and we refer to Hassin and Haviv (2003) and Hassin (2016) for a comprehensive review. In more recent work, Hassin and Roet-Green (2017) and Hu et al. (2018) explore information heterogeneity and find that providing queue-length information to a fraction of customers but not all may improve system performance. Wang et al. (2019) studies an M/M/1 priority queue with balking under observable/unobservable settings.

The effects of different levels of delay information, with different degrees of precision, on the overall system was examined in Guo and Zipkin (2007). Veeraraghavan and Debo (2009) explores how the queue-length information could be used by customers to infer service quality. Using a cheap talk model, Allon et al. (2011) examines how providing nonverifiable delay information can improve firm's profits and customer utility. Hassin and Roet-Green (2018) considers a setting where customers must travel to the queue to

get served. They study the effect of providing queue information to customers before they decide whether to balk or travel to the store. Our paper connects to this literature by studying the implications of firm's information disclosure policy where orders could be processed in parallel while customer travels to the queue. We refer to Ibrahim (2018) for a comprehensive review on sharing delay information in service systems.

A growing literature on omnichannel retail examines impact of different omnichannel strategies and fulfilment methods, both analytically and empirically (e.g., Gao and Su, 2017a; Gallino and Moreno, 2014; Nageswaran et al., 2020). The focus of the work in omnichannel retail is on considerations such as inventory and product management, which are different from the considerations that are crucial in the context of omnichannel services, such as throughput and consumer surplus. The common thread in omnichannel management is exploring the effects of committing inventory or capacity to customers before they arrive at the retail outlet. Similar issues arise in the context of restaurant reservations (Cil and Lariviere, 2013), medical appointment scheduling (Ahmadi-Javid et al., 2017), and delivery in food services (Feldman et al., 2019; Chen et al., 2019). Our paper contributes to the growing but scant literature in omnichannel services.

Gao and Su (2017b) and Kang et al. (2020) adopt a two-stage tandem-queue (stage 1 is cashier queue and stage 2 is food preparation queue) to model omnichannel system. The online customers, in their model, bypass the first stage queue by self-ordering, whereas the walk-in customers go through both stages to get served. Gao and Su (2017b) considers the capacity planning problem for firms adopting self-ordering technology. Contrary to popular belief that self-ordering technology would replace human workers, they find

that it is sometimes optimal for firms to increase workforce level. Kang et al. (2020) focuses on exploring the importance of prioritization design choices on omnichannel systems in terms of the throughput and social welfare. They show that implementing a wrong prioritization policy can have detrimental effects on the throughput of an omnichannel system with impatient customers. These two papers differ from the rest of the literature on omnichannel services, including our paper, in that their primary focus is on tactical queue management.

Baron et al. (2019) explores customer channel choices in omnichannel systems. In their model, all customers make a one-shot decision of whether to order online or to walk-in. Customers who choose to order online, do not possess any information about the state of the queue whereas customers who walk-in can observe the number of orders from both channels. They find that, when online ordering option is available, individual customer utility and social welfare could be reduced because of intra-channel interference by online orders on the walk-in channel.

Roet-Green and Yuan (2019) analyzes a partially observable omnichannel system where online customers are invisible, and the store customers are visible; all customers have the same information structure. They show that a two-channel (partially observable) system is more profitable for the firm if customers are not deterred from joining, by the presence of an invisible queue. However, if the customers are indeed deterred from joining, it could make a two-channel (partially observable) system more socially desirable compared to a single channel (fully observable) system. They also consider the case where service of the visible class is prioritized.

Liu and Yang (2020) explores the effect of providing queue information to remote customers, on the firm's throughput. They compare two models: order-ahead (in which all customers who wants to place an order must do so using the online channel before they travel to the store for pick-up), and order-onsite (in which customers can only order in the physical store). They consider various information provision policies and shows that under full information provision, their order-ahead model achieves higher throughput than order onsite model. However, if information is not provided to remote customers, then their order-onsite model could yield a higher throughput.

Our paper best relates to Baron et al. (2019), Liu and Yang (2020) and Roet-Green and Yuan (2019). A primary focus of our paper is to characterize customer channel choice behavior and study how it is affected by queue information disclosure when customers travel to the store. To the best of our knowledge, Baron et al. (2019) is the only other paper that considers the customer channel choice problem in the existing literature on omnichannel services. Our work is differentiated in that the offline ordering option in our model is available to app users as well; app users could either order online or defer making an ordering decision until they arrive at the store and observe the queue. We capture the temporal evolution of the queue as app users travel to the store and study how it factors in their channel choice strategy. Moreover, we model quality sensitivity, and consider two customer types with potentially different impatience levels. In terms of the information structure, our paper relates to Liu and Yang (2020) in that the visibility for app-users depends on the firm's information disclosure policy. On the other hand, non-app users, observe orders from both channels (similar to Baron et al., 2019). With information disclosure, analyzing the evolution of the queue while the app user travels,

is intractable under a continuous-time model. A novel feature that separates our paper from the existing literature on omnichannel services is that we adopt a discrete-time queuing model to tackle this intractability. Our results compliment Liu and Yang (2020), Roet-Green and Yuan (2019), and Baron et al. (2019) by showing that, depending on the relative impatience of app users and non-app users, both of our omnichannel models (with and without information) could deliver a lower throughput than a single channel system. Our contribution in terms of the core message is that the non-app users are disadvantaged by the online orders and their welfare, relative to the app users, needs to be seriously considered for a proper evaluation of the overall performance of omnichannel systems.

2.2. Model

In this section, we describe our model. We first describe customer types, their associated costs and other parameters of the service system. Following this, we will present the information structure in the model and define a sequence of events. We also examine the expected utility of app users.

2.2.1. Customers, Costs and the Service System

We model strategic interaction of customers in *omnichannel* service systems, which provide customers the option of ordering *online* (for e.g. using mobile apps), in addition to the conventional *offline* option of ordering in the physical store. We consider two types of customers: i) *app users* who are tech-savvy and may choose between either the online or the offline ordering options, and ii) *non-app users* who never use the online option. All

customers are utility maximizers, and have the same valuation, $v > 0$ for the service. We further assume that all customers (regardless of their channel choice) pay the same price for service.

We say that an app user *arrives at the market* when service requirement arises and she is located away from the physical store. An app user arriving at the market may choose to order online, in which case the order is immediately placed in the queue. She then travels to the physical store to pick it up, with the possibility that her order might be ready by the time she reaches the store. Customers prefer to have their product as soon as they are prepared, and in this sense they are quality sensitive (i.e. no one wants a lukewarm latte). If an app user's order gets prepared before she reaches the store, she incurs a fixed cost, $c_q \geq 0$, for degraded product quality. This results in her utility being, $v - c_q$. Since non-app users always walk-in, they never incur any quality cost.

Alternatively, an app user may choose to not order online, travel to the store and then decide whether or not to order after observing the queue. In this case, there is a possibility that the system might be congested, which would result in her waiting in the store before she could place her order. Both sets of customers are sensitive to the amount of time they spend waiting in the store. The wait sensitivities for app users and non-app users are denoted by $c_{wT} \geq 0$ and $c_{wN} \geq 0$ respectively (T denotes tech-savvy and N denotes non-tech-savvy). Thus, if an app user or a non-app user has to wait in the store for the completion of k orders, including her own, then her resultant utility would be $v - \mathbb{E}W(k) \cdot c_{wT}$ or $v - \mathbb{E}W(k) \cdot c_{wN}$ respectively, where $\mathbb{E}W(k)$ denotes the expected number of time periods until she's served. We assume that there is no cost associated with a customer traveling to the physical store (see §2.6 for a discussion). As a results,

app users who choose the offline option always travel to the store and see the queue before they decide whether to seek service or to balk.

We use a single discrete-time queue (see Remark 2) that serves customers in a first-come-first-served (FCFS) manner. At each time period t, S_t denotes the number of service slots that are generated, where S_t is an independent and identically distributed random variable with mean μ. Every customer who joins the queue needs one service slot. We are agnostic about the number of servers generating these slots, which is independent of the mix of customers in the queue and is statistically identical each period. If the number of orders in the system is smaller than the number of generated service slots, the excess service slots get wasted. Thus, the number of orders getting processed in period t, would be the minimum of the number of service slots generated and the number of orders in the system, and in this sense, μ denotes the service *capacity* of the system. At any point in time, length of the queue denotes the total number of orders in the system that are waiting to be processed. If S_t is a Geometric random variable with expected value μ, then $\mathbb{E}W(k) = \frac{k}{\mu}$ (see Appendix B.1), where $\mathbb{E}W(k)$ is defined in the preceding paragraph. In general $\mathbb{E}W(k)$ is increasing in k and decreasing in μ. In the rest of the paper, we are going to approximate $\mathbb{E}W(k)$ by $\frac{k}{\mu}$ even for general distributions of S_t, which yields good structural insights.

2.2.2. Information Structures and Sequence of Events

We make a simplifying assumption that the entire order-queue (which includes both offline and online orders) is visible to all customers who walk-in to the store (see Baron et al., 2019). For the app users who arrive at the market, we consider the two following settings:

- **Model with Information.** The firm provides real-time queue-length information (consistent with the literature, see Ibrahim, 2018) through the app. Thus, the app users know the state of the system before they decide whether to order online or to defer the ordering decision and re-sample the queue upon arrival at the store.
- **Model without Information.** The firm does not provide any information about the system, and thus the app users make decisions based on their steady-state beliefs about the system's state. They decide whether to order early but blindly, or choose the offline option to sample the queue.

Without loss of generality, we assume that at every time period exactly one app user arrives at the market. We denote the arrival rate of app users by Λ_T, and thus we have $\Lambda_T = 1$. Furthermore, it takes exactly one time period for an app user to travel from the market to the store (see Remark 2). At each time period t, a sequence of four events (see Figure 2.1) take place in the following order:

(1) *Event 1: Arrival of Non-App Users.* At the beginning of time period t, non-app users arrive at the physical store. The number of non-app users who arrive at the store is denoted by A_t, which is an independent and identically distributed random variable with mean Λ_N. We specify the probabilities of arrivals as $\mathbb{P}(A_t = k) = a_k$. The length of the queue in period t before the arrival of the non-app users is denoted by $x_0(t)$. Since customers are wait-sensitive, not everyone who arrives at the store, joins the queue upon observing $x_0(t)$. The effective number of non-app users who join the system in period t is denoted by n_t. We assume

that the outside option for a customer who balks, is zero. We denote the updated queue length after the arrival of non-app users by $x_1(t) = x_0(t) + n_t$.

(2) *Event 2: Arrival of App User at the Store.* Next, the app user who arrived at the market in time period $t-1$, arrives at the physical store (it takes one time period after arrival at the market to arrive at the store), for either of the following two purposes:

- *To pick up her order, if she ordered online in period* $t-1$: If her order was prepared in period $t-1$, she incurs a quality cost and leaves the system immediately without incurring any waiting cost. Otherwise, she does not incur any quality cost, but does incur a cost of waiting in the store, for her order to get processed. Since orders are processed in an FCFS manner, she has to wait until all the orders placed before hers, as well as her own, gets processed. Position of her order in the queue, when she arrives at the store, is $x_0(t)$. Thus, she has to wait for the completion of $x_0(t)$ orders before she leaves. We note that, the joining of non-app users in period t, does not affect the position of her order, and hence, does not affect her waiting cost. We denote the queue length after this event by $x_2(t)$. Since, the app user already placed her order in period $t-1$, this event has no bearing on the

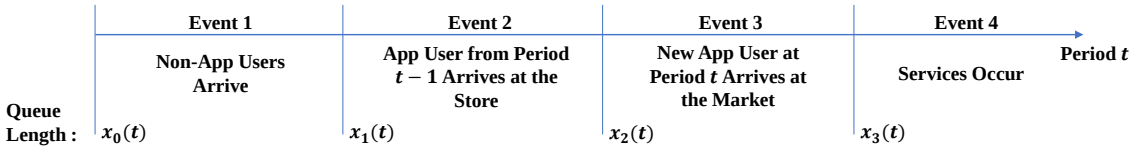


Figure 2.1. Sequence of Events in Time Period t

evolution of the state of the system in period t, and hence, the queue length remains unchanged after this event, i.e. $x_2(t) = x_1(t)$.

- *To decide whether to place an order, if she did not order online in period* $t-1$: She observes the current queue length $x_1(t)$, and decides whether or not to order. After this event, the queue length is updated to $x_2(t) = x_1(t) + 1$, if the app user joins, otherwise the queue length remains unchanged, i.e., $x_2(t) = x_1(t)$. Balking yields an utility of zero.

(3) *Event 3: Arrival of App User at the Market.* The next event in time period t is the arrival of a new app user at the market. For the *model with information*, the current state of the queue, $x_2(t)$, is revealed to this app user. She takes this information into account to decide whether to order online immediately, or to defer her decision of whether to order or balk, until she arrives at the store in period $t + 1$ (for brevity, we will henceforth refer to this action as *choosing the offline option*). For the *model without information*, the app user's decision is based on her rational belief over the steady state queue length, $x_2(t)$. The queue length following this event is updated as $x_3(t) = x_2(t) + 1$ if the app user orders online, otherwise, it remains unchanged, i.e., $x_3(t) = x_2(t)$.

(4) *Event 4: Services.* Finally, service slots are generated at the end of time period t. The probability of the number of service slots generated is denoted by $\mathbb{P}(S_t = k) = s_k$. The queue length following this event is identical to the queue length at the start of time period $t + 1$, and is given by $x_0(t + 1) = (x_3(t) - S_t)^+$, where z^+ denotes $\max(0, z)$.

2.2.3. Customer Utility of App Users

In the model with information, upon arrival at the market, an app user can see the state of the queue, $x_2(t)$, before deciding whether or not to order online. If she observes $x_2(t) = L$ and orders online, her expected utility is given by,

$$U_o(L) = v - \underbrace{\mathbb{E}_{S_t}[\mathbb{1}(S_t \geq L+1)] \cdot c_q}_{\text{Expected quality cost}} - \underbrace{\mathbb{E}_{S_t}[(L+1-S_t)^+] \cdot \frac{c_{wT}}{\mu}}_{\text{Expected waiting cost}}. \tag{2.1}$$

where $\mathbb{1}(\cdot)$ denotes the indicator function. She incurs a quality cost if the number of service slots, S_t, generated in period t, is greater than or equal to the number of orders (queue length) currently in the system, in addition to her own order. Otherwise, she incurs waiting cost if there are leftover orders to be processed in period $t+1$. The term $(L+1-S_t)^+$ is equivalent to $x_0(t+1)$ which is the queue length at the beginning of time period $t+1$. $x_0(t+1)$ also denotes the position of her order in the queue, in the event that her order is not processed before her arrival.

Upon observing $x_2(t) = L$, her expected utility of choosing the offline option is,

$$U_s(L) = \mathbb{E}_{S_t, A_{t+1}}[\max(0, \hat{U}_s(L))], \tag{2.2}$$

where

$$\hat{U}_s(L) = v - \Big((L - S_t)^+ + n_{t+1}(L) + 1\Big) \cdot \frac{c_{wT}}{\mu} \tag{2.3}$$

represents the utility associated with the action of going to the store and placing an order. We note that in (2.2), the first argument within the max function is zero. This accounts

for the fact that, an app user who does not order online in period t, can potentially balk at the store, upon observing the queue length $x_1(t+1)$ in period $t+1$. The term $(L - S_t)^+ + n_{t+1}(L)$ in (2.3) is equivalent to $x_1(t+1)$. Moreover, $n_{t+1}(L)$ denotes the dependence of the number of arrival of non-app users in period $t+1$, on L, the observed queue length by an app user arriving at the market in period t.

If queue information is unavailable to app users, then their utilities of ordering online, and choosing the offline option are given by,

(2.4)

$$\mathbb{E}_{x_2}(U_o(x_2)) = \sum_{L=0}^{\infty} \mathbb{P}(x_2 = L) \cdot U_o(L) \qquad \text{and} \qquad \mathbb{E}_{x_2}(U_s(x_2)) = \sum_{L=0}^{\infty} \mathbb{P}(x_2 = L) \cdot U_s(L)$$

respectively. We note that these expected utilities depend on the steady state distribution of the queue. In order to denote the steady state, we suppress the time argument in $x_2(t)$.

Remark 2. *An app user arriving at the market is uncertain of when her order will be ready if she orders early but is also uncertain of how long the line will be if she chooses to visit the store. Probabilistically describing how the system will evolve is much simpler in a discrete-time model than in a continuous-time model. We use a discrete-time model in order to maintain tractability while capturing important features of the setting. Additionally, this provides a fair amount of flexibility in the way we model services and arrivals. Moreover, in regards to our assumption, that it takes one time period for an app user to travel from the market to the store, the queue can still change significantly, and is driven by both the services and the arrival of non-app users. Although this happens in a way that is fundamentally limited, it maintains tractability.*

2.3. Customer Strategy

In this section, we characterize the ordering strategies for non-app users and app users under general arrival and service processes for the models with and without information. For the proofs of Theorem and Proposition in this section, we refer to Appendix B.4.

2.3.1. Store Customers

We refer to the non-app users and the app users who choose the offline option as store customers. Upon observing the queue, store customers decide whether to join the queue or balk. We note that app users who choose to order online, arrive at the store only for pick-up and do not make any join/balk decision. Using an approach akin to Naor (1969), it is straightforward to show that store customers join only if the observed queue length is below a threshold, and they balk otherwise.

Non-app users, upon observing queue length $x_0(t)$, join only if their expected utility is non-negative. We assume that arriving non-app users are randomly ordered and make joining decision sequentially. Let n_t denote the number of non-app users who join the queue in period t. Then we have $v - \frac{x_0(t)+n_t}{\mu} c_{wN} \geq 0$, which implies

$$n_t \leq \left(\left\lfloor \frac{v\mu}{c_{wN}} \right\rfloor - x_0(t) \right)^+ \tag{2.5}$$

where $\lfloor z \rfloor$ denotes the greatest integer less than or equal to z. Thus, non-app users join only if the observed queue length, $x_0(t)$, is strictly less than the threshold $\left\lfloor \frac{v\mu}{c_{wN}} \right\rfloor$, and balk otherwise. Hence, the effective number of non-app users joining the queue in period t, is given by $n_t = \min(A_t, (\lfloor \frac{v\mu}{c_{wN}} \rfloor - x_0(t))^+)$.

An app user who chooses the offline option arrives at the store and observes queue length $x_1(t)$. Similar to non-app users, this app user joins the queue only if her expected utility is non-negative, i.e. $v - \frac{x_1(t)+1}{\mu} c_{wT} \geq 0$. This implies that the app user joins only if

$$x_1(t) \leq \left(\left\lfloor \frac{v\mu}{c_{wT}} \right\rfloor - 1 \right)^+, \tag{2.6}$$

and balks, otherwise.

We denote the store threshold for the app users and non-app users by $\tau_s = \left\lfloor \frac{v\mu}{c_{wT}} \right\rfloor$ and $\tau_n = \left\lfloor \frac{v\mu}{c_{wN}} \right\rfloor$ respectively. To avoid trivialities, we generally assume that $\tau_s \geq 1$ and $\tau_n \geq 1$, such that store customers are willing to join, upon observing an empty queue. Both τ_n and τ_s are similar to *Naor's threshold* for observable queues (see Naor, 1969).

2.3.2. Online Customers: Model with Information

If queue information is available, an app user arriving at the market in period t faces no uncertainty regarding the ordering decision of any app user who arrived before period t. The state of the queue, $x_2(t)$, which is observable to the new app user, captures all the past app user decisions. Thus, this app user follows a *dominant strategy* in which she orders online if $U_o(L) > U_s(L)$, and chooses the offline option, if $U_o(L) \leq U_s(L)$. Before characterizing app users' channel choice strategy, we make the following assumption.

Assumption 2. *We assume that* $\mathbb{P}(S_t = L) > 0$ *and the function* $\mathcal{D}(L)$ *given by,*

$$\mathcal{D}(L) := \frac{\left(\mathbb{E}_{S_t, A_{t+1}}[n_{t+1}(L)] \cdot \frac{c_{wT}}{\mu} - g(L) \right) - \left(\mathbb{E}_{S_t, A_{t+1}}[n_{t+1}(L-1)] \cdot \frac{c_{wT}}{\mu} - g(L-1) \right)}{\mathbb{P}(S_t = L)},$$

is decreasing for all non-negative integer L*, where* $g(L) = U_s(L) - \mathbb{E}_{S_t, A_{t+1}}(\hat{U}_s(L))$.

Now, we characterize the channel choice strategy for app users in the proposition below.

Proposition 4. (Channel Choice Strategy)

(i) If $\frac{c_{wT}}{\mu} - c_q \geq 0$*, then there exists a threshold* $\tau_u \geq 0$ *such that an app user arriving at the market orders online if* $0 \leq x_2(t) \leq \tau_u$*, and chooses the offline option, otherwise.*

(ii) If $\frac{c_{wT}}{\mu} - c_q < 0$ *and Assumption 2 holds, then*

(a) If $U_o(L) \geq U_s(L)$ *for some non-negative integer* L*, there exist two thresholds* $\tau_l \geq 0$ *and* $\tau_u \geq 0$*, such that, the app user orders online if* $\tau_l \leq x_2(t) \leq \tau_u$*, and chooses the offline option, otherwise.*

(b) If $U_o(L) < U_s(L)$ *for all* $L \geq 0$*, the app user always chooses the offline option.*

If quality sensitivity is sufficiently low, then upon arriving at the market and observing a long line, an app user chooses the offline option. Intuitively, a long line promises a long wait in the store and an app user delays ordering to see if there is a favorable evolution of the queue, in which case the app user orders at the store, and otherwise balks. On the other hand, if quality sensitivity is relatively high compared to the waiting cost, then if she observes a short queue, there's a high likelihood that her order would be prepared prior to her arrival at the store. Thus, exploiting the opportunity to sample the queue twice, app users follow a *dual-threshold* policy where they alternate between walking in (to avoid the quality penalty), ordering online (to shorten their wait), and walking in (to avoid a very long wait) as the queue grows. If the quality sensitivity is extremely high

compared to the wait sensitivity, for e.g., if $c_{wT} = 0$ and $c_q > 0$, then the online channel is not viable, i.e. $U_o(L) < U_s(L)$ for all $L \geq 0$, and app users always choose the offline option.

2.3.3. Online Customers: Model without Information

In the absence of queue information, app users' channel choice strategy depends on their rational belief over the steady state queue length, x_2. The steady state, in turn, depends on the strategy followed by the app users. Thus, an equilibrium emerges, as presented in Theorem 4. Since app users are homogeneous, we have a symmetric equilibrium strategy, θ, which represents the probability than an app user orders online. The expected utilities of ordering online and choosing the offline option are given by,

$$\overline{U}_o(\theta) = \mathbb{E}_{x_2}(U_o(x_2)) \qquad \text{and} \qquad \overline{U}_s(\theta) = \mathbb{E}_{x_2}(U_s(x_2)) \tag{2.7}$$

respectively, where $U_o(x_2)$ and $U_s(x_2)$ are given by (2.1) and (2.2).

Theorem 4. (Existence of Equilibrium) *A symmetric Nash equilibrium, θ, for app users arriving at the market in an omnichannel system without information, exists and is given as follows:*

(i) If $\overline{U}_o(1) > \overline{U}_s(1)$ then $\theta = 1$.

(ii) If $\overline{U}_o(0) < \overline{U}_s(0)$ then $\theta = 0$.

(iii) If $\overline{U}_o(1) \leq \overline{U}_s(1)$ and $\overline{U}_o(0) \geq \overline{U}_s(0)$ then there exists a $\theta^* \in [0, 1]$ such that $\overline{U}_o(\theta^*) = \overline{U}_s(\theta^*)$ and $\theta = \theta^*$.

In Theorem 4(iii), although we do not prove uniqueness, through all the computational examples we have studied, we never encounter multiple equilibria.

Quality sensitivity is a key driver in determining when app users switch from ordering online to choosing the offline option. We recall from (2.1), that $U_o(L)$ is monotonically decreasing in c_q. Consequently, it is straightforward to show that, if all app users order online, i.e. $\theta = 1$, for a given set of parameters with some quality sensitivity $\hat{c}_q$, then the same equilibrium $\theta = 1$ continues to hold for all $c_q < \hat{c}_q$. Conversely, if all app users choose the offline option, i.e. $\theta = 0$, for a set of parameters with some quality sensitivity $\hat{c}_q$, all app users continue to choose the offline option, i.e. $\theta = 0$, for all $c_q > \hat{c}_q$.

2.4. Analytical Study

In this section, we analytically characterize the average throughput and the average consumer surplus for single channel and omnichannel systems. To maintain tractability we consider systems with the following assumptions:

- The number of arrivals of non-app users is such that $\mathbb{P}(A_t = 0) = \frac{1}{2}$ and $\mathbb{P}(A_t = 1) = \frac{1}{2}$.
- The number of service slots generated is such that $\mathbb{P}(S_t = 1) = \frac{1}{2}$ and $\mathbb{P}(S_t = 2) = \frac{1}{2}$.
- Wait-sensitivity of non-app users is such that $\frac{3}{4} < \frac{c_{wN}}{v} \leq \frac{3}{2}$, which implies that non-app users join the system only if the queue is empty, i.e. $\tau_n = 1$. We consider the following two scenarios for the wait-sensitivity of app users:
  (1) The wait-sensitivity of app users is comparable to the non-app users, i.e., we assume that $\frac{3}{4} < \frac{c_{wT}}{v} \leq \frac{3}{2}$. This implies that the app users at the store

join only if the queue is empty, i.e. $\tau_s = 1$. Alluding to the app users' wait-sensitivity, throughout the rest of this section we will refer to this setting as the *Impatient* scenario.

(2) App users are less wait-sensitive than the non-app users, i.e., we assume that $\frac{1}{2} < \frac{c_w T}{v} \leq \frac{3}{4}$. This implies that the store threshold for app users is $\tau_s = 2$. Since, in this setting, the app users are more patient, throughout the rest of this section we will refer to this as the *Patient* scenario.

We note that, these assumptions result in a smaller state space, which allows us to characterize the steady state of the queue. This also allows us to explicitly solve for the online ordering thresholds τ_l and τ_u without imposing Assumption 2. The qualitative aspects of the results presented in this section extend to systems with more general arrival and service processes for a wide range of parameters and customer primitives, which will be illustrated in §2.5. For the details on the steady state calculations we refer to Appendix B.2. For the proofs of results in this section we refer to Appendix B.5.

2.4.1. Customer Strategy and Throughput

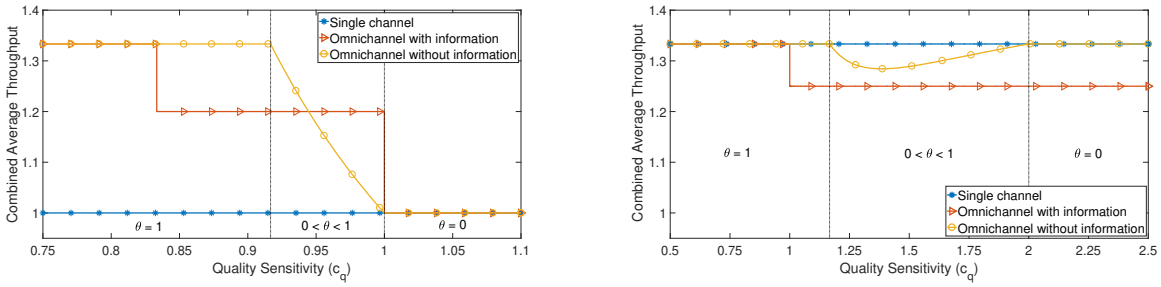
In this section, we present and compare the customer strategy and the average throughput for both scenarios, Impatient and Patient. First, we focus on customer strategy.

Proposition 5. *The probability, θ, that an app user arriving at the market orders online, in the omnichannel (without information) system for the Impatient scenario, is*

given by

$$\theta = \begin{cases} 1 & \text{if } 0 < \dfrac{c_q}{c_{wT}} \leq \dfrac{1}{6} + \dfrac{3}{4} \cdot \dfrac{v}{c_{wT}}, \\ \dfrac{12(v - c_q)}{3v - 2c_{wT}} & \text{if } \dfrac{1}{6} + \dfrac{3}{4} \cdot \dfrac{v}{c_{wT}} < \dfrac{c_q}{c_{wT}} < \dfrac{v}{c_{wT}}, \\ 0 & \text{if } \dfrac{v}{c_{wT}} \leq \dfrac{c_q}{c_{wT}} < \infty. \end{cases}$$

Figure 2.2a illustrates the ordering strategy, θ, for app users in the Impatient scenario as a function of their quality sensitivity. From Proposition 5 we note that, when quality sensitivity is dominated by wait sensitivity, most app users order online. By ordering online, app users reduce the likelihood of having to wait in the store for their order. Keeping all else fixed, with increasing quality sensitivity, app users become more likely to not order online and instead choose the offline option to avoid potential quality degradation of the product. Now, in an omnichannel system with queue information, app users' strategy would depend on the queue length they observe. From Table 2.1 we note that, as the



(a) Average Throughput for Impatient Scenario

(b) Average Throughput for Patient Scenario

Figure 2.2. Comparision of combined (app users and non-app users) average per period throughput for the Impatient and Patient Scenarios, corresponding to single channel, omnichannel (with information) and omnichannel (without information) systems. The graph also illustrates the app users' ordering strategy, θ, for the omnichannel (without information) system.

Ordering strategy when			
$x_2(t)$	$0 < \frac{c_q}{c_{wT}} < \frac{1}{3} + \frac{1}{2} \cdot \frac{v}{c_{wT}}$	$\frac{1}{3} + \frac{1}{2} \cdot \frac{v}{c_{wT}} \leq \frac{c_q}{c_{wT}} < \frac{v}{c_{wT}}$	$\frac{v}{c_{wT}} \leq \frac{c_q}{c_{wT}} < \infty$
0	Online	Offline	Offline
1	Online	Online	Offline

Table 2.1. Ordering strategy for app users arriving at the market, for the omnichannel (with information) setting in the Impatient scenario, as a function of the queue length, $x_2(t)$.

ratio of quality sensitivity to wait sensitivity increases, all app users eventually choose the offline option. This is similar to the case without information. Thus, both omnichannel systems behave as a single channel system as quality sensitivity increases.

We, next, contrast this with app users' strategy for the Patient scenario.

Proposition 6. *The probability, θ, that an app user arriving at the market orders online, in the omnichannel (without information) system for the Patient scenario, is given by*

$$\theta = \begin{cases} 1 & \text{if } 0 < \frac{c_q}{c_{wT}} \leq \frac{7}{6}, \\ \frac{6c_{wT} - 3c_q}{3c_q - c_{wT}} & \text{if } \frac{7}{6} < \frac{c_q}{c_{wT}} < 2, \\ 0 & \text{if } 2 < \frac{c_q}{c_{wT}} < \infty. \end{cases}$$

Similar to the Impatient scenario, in the Patient scenario, the probability that an app user orders online, θ, decreases with increasing quality sensitivity of app users. Thus, for a high enough ratio of quality sensitivity to wait sensitivity, all app users choose the offline option. However, in this regard, the ordering strategy for app users in the omnichannel (with information) setting for the Patient scenario (see Table 2.2) is different from the

Impatient scenario. No matter how high the ratio of quality sensitivity to wait sensitivity is, app users always order online if they observe a queue length of 2, upon arriving at the market. The app user's order would not get processed in the current time period since only a maximum of 2 orders can be processed per period in this system. Thus, no quality cost is incurred in this case. This illustrates how various aspects of the arrival and service processes, and the evolution of the queue when the app users travel from the market to the store, factor in app users' ordering decision.

Furthermore, from Proposition 6 and Table 2.2, we observe that, keeping quality sensitivity fixed, app users are more likely to order online as they become more wait-sensitive. By ordering online, app users reduce the likelihood of having to wait in the store for their order. Note that the ratio of valuation over wait sensitivities are constrained in the Patient and Impatient scenarios. Under a more general setting in §2.5, we will illustrate that, as the wait sensitivity for app users increases keeping all else fixed, all app users arriving at the market eventually choose the offline option. Intuitively, for moderate wait sensitivities, app users prefer ordering in advance to lower the likelihood of waiting in the store. Highly wait-sensitive app users, on the other hand, prefers delaying making

Ordering strategy when			
$x_2(t)$	$0 < \frac{c_q}{c_{wT}} < 1$	$1 \leq \frac{c_q}{c_{wT}} < \frac{4}{3}$	$\frac{4}{3} \leq \frac{c_q}{c_{wT}} < \infty$
0	Online	Offline	Offline
1	Online	Online	Offline
2	Online	Online	Online

Table 2.2. Ordering strategy for app users arriving at the market, for the omnichannel (with information) setting in the Patient scenario, as a function of the queue length, $x_2(t)$.

an ordering decision until reaching the store for a potentially favorable evolution of the queue. Thus, the ordering strategy for app users, θ, is not necessarily monotone in app users' wait sensitivity.

Next, we discuss the comparison of average throughput.

Proposition 7. *The combined average per period throughput for the Impatient scenario, corresponding to single and omnichannel systems are given as follows:*

(1) Omnichannel without information:

$$\lambda^{ONI} = \begin{cases} \frac{4}{3} & \text{if } 0 < \frac{c_q}{c_{wT}} \leq \frac{1}{6} + \frac{3}{4} \cdot \frac{v}{c_{wT}}, \\ \frac{3v - 2c_{wT}}{3c_q - 2c_{wT}} & \text{if } \frac{1}{6} + \frac{3}{4} \cdot \frac{v}{c_{wT}} < \frac{c_q}{c_{wT}} < \frac{v}{c_{wT}}, \\ 1 & \text{if } \frac{v}{c_{wT}} \leq \frac{c_q}{c_{wT}} < \infty, \end{cases}$$

(2) Omnichannel with information:

$$\lambda^{OI} = \begin{cases} \frac{4}{3} & \text{if } 0 < \frac{c_q}{c_{wT}} < \frac{1}{3} + \frac{1}{2} \cdot \frac{v}{c_{wT}}, \\ \frac{6}{5} & \text{if } \frac{1}{3} + \frac{1}{2} \cdot \frac{v}{c_{wT}} \leq \frac{c_q}{c_{wT}} < \frac{v}{c_{wT}}, \\ 1 & \text{if } \frac{v}{c_{wT}} \leq \frac{c_q}{c_{wT}} < \infty, \end{cases}$$

(3) Single channel:

$$\lambda^S = 1.$$

See Figure 2.2a for illustration of Proposition 7. First, note that the average throughput for omnichannel system with information is a step-function where each jump corresponds to a change in the online ordering thresholds, τ_l and τ_u. In contrast, the average throughput for omnichannel system without information changes smoothly with θ.

Figure 2.2a illustrates that, in the Impatient scenario, the omnichannel systems result in higher average throughput compared to single channel. The online ordering option draws in the demand early, which results in higher app user throughput. Non-app users, on the other hand, get crowded out by the presence of the online channel. In the Impatient scenario, going from single channel to omnichannel, increase in app user throughput outweighs the reduction in non-app user throughput which results in increased overall throughput.

Without information disclosure, channel choice made by app users arriving at the market is based on their rational beliefs about the steady state of the system. If quality sensitivity is low, then most app users would prefer ordering in advance. By disclosing queue information, these customers would be deterred from ordering online in those unfavorable states where queue length is too short (see Table 2.1). A shorter queue promises a quicker service completion and hence a higher likelihood of incurring a quality penalty. This lowers the average throughput for the omnichannel system with information. On the other hand, if quality sensitivity is high, most app users would choose the offline option in the absence of information. Disclosing queue information to these customers, in this case, would lure them into placing online orders in favorable states of the system. This increases the average throughput for the omnichannel system with information. Thus,

neither of the two omnichannel systems delivers the highest average throughput across all parameter regimes.

Now, we compare these results with the average throughput for the Patient scenario.

Proposition 8. *The combined average per period throughput for the Patient scenario corresponding to single and omnichannel systems are given as follows:*

(1) Omnichannel without information:

$$\lambda^{ONI} = \begin{cases} \frac{4}{3} & \text{if } 0 < \frac{c_q}{c_{wT}} \leq \frac{7}{6}, \\ 1 + \frac{18c_q^2 - 36c_q c_{wT} + 20c_{wT}^2}{36c_q^2 - 51c_q c_{wT} + 18c_{wT}^2} & \text{if } \frac{7}{6} < \frac{c_q}{c_{wT}} < 2, \\ \frac{4}{3} & \text{if } 2 \leq \frac{c_q}{c_{wT}} < \infty, \end{cases}$$

(2) Omnichannel with information:

$$\lambda^{OI} = \begin{cases} \frac{4}{3} & \text{if } 0 < \frac{c_q}{c_{wT}} < 1, \\ \frac{5}{4} & \text{if } 1 \leq \frac{c_q}{c_{wT}} < \frac{4}{3}, \\ \frac{5}{4} & \text{if } \frac{4}{3} \leq \frac{c_q}{c_{wT}} < \infty, \end{cases}$$

(3) Single channel:

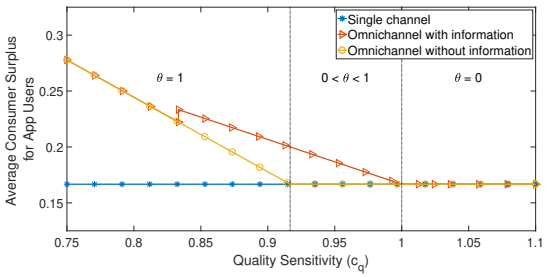
$$\lambda^S = \frac{4}{3}.$$

Surprisingly, Figure 2.2b illustrates that, the single channel system outperforms the omnichannel systems in terms of average throughput in the Patient scenario. Note that

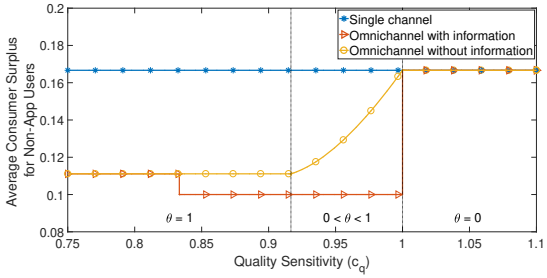
the non-app users are more wait-sensitive compared to the app users. As a result, in the omnichannel systems, enough non-app users get crowded out by the app users that the overall throughput drops in comparison to the single channel setting.

Furthermore, we observe from Figure 2.2b that the average throughput for omnichannel (without information) setting drops when app users follow a mixed equilibrium strategy, i.e. they randomize between the online and the offline ordering options. When app users either order online, i.e. $\theta = 1$, or choose the offline option, i.e. $\theta = 0$, their arrival process is deterministic since we have assumed that only one app user arrives every time period. In contrast, when app users follow a randomized strategy, i.e. $0 < \theta < 1$, their arrival process becomes stochastic. Due to increase in variability in the arrival process, the variability in queue length increases as well. Due to this, the queue observed by the non-app users becomes stochastically larger when $0 < \theta < 1$. This, in turn, results in more non-app users to balk in the omnichannel setting compared to the single channel setting.

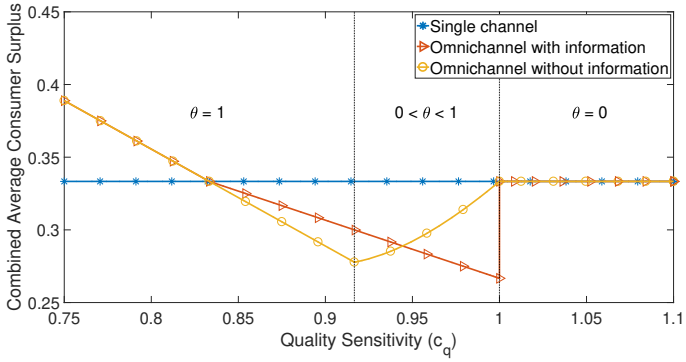
Similarly, for the omnichannel (with information) setting, the average throughput drops in comparison to single channel. This happens when the ordering strategy for app users (see Table 2.2), is such that, depending on the observed queue length, app users go back and forth between the online and choosing the offline options. Thus, omnichannel systems might not necessarily outperform a single channel system in terms of delivering the highest average throughput. In §2.5, we show that these results hold under a more general system with a wide range of system parameters and customer primitives.



(a) Average Consumer Surplus for App Users



(b) Average Consumer Surplus for Non-App Users



(c) Combined Average Consumer Surplus (App Users and Non-App Users)

Figure 2.3. Comparison of average per period consumer surplus in the Impatient scenario for single channel, omnichannel (with information) and omnichannel (without information) systems. For illlustrative purposes, valuation, app user wait sensitivity and non-app user wait sensitivity are fixed at $v = 1$, $c_{wT} = 1$ and $c_{wN} = 1$.

2.4.2. Consumer Surplus

In this section, we present the analytical results on consumer surplus. We focus on the Impatient scenario. In characterizing the average per period consumer surplus, we only consider app users and non-app users who get served (outside option associated with balking has zero utility). For ease of exposition, we use Figure 2.3 to illustrate these

results and we refer to Lemmas 17, 19 & 21 in Appendix B.5 for the exact analytical expressions.

Moving from a single channel to omnichannel systems, the presence of app users in the system diminishes the consumer surplus for non-app users. From Figure 2.3 (a) and 2.3 (b) we observe that, for low to moderate quality sensitivities, app users benefit from ordering online in the omnichannel systems. In doing so, the non-app users get crowded out. In §2.5 we will illustrate that non-app users' consumer surplus is consistently higher in the single channel setting, for a wide range of parameters. We will also show that, in contrast with the non-app users, the app users might not always prefer an omnichannel system.

We observe from Figure 2.3 (a) that the app users benefit from queue information disclosure. However, the effect of queue information disclosure on the overall system depends on the combined effect on both customer types (see Figure 2.3 (c)).

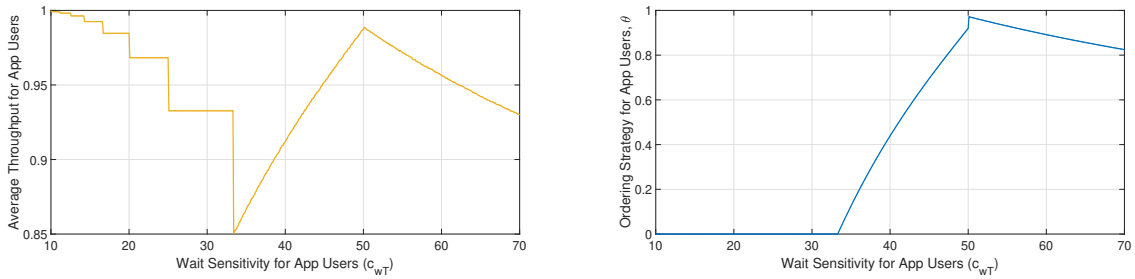
In the Patient scenario, patient app users crowd out less patient non-app users. The omnichannel systems serve more app users, and deliver higher combined consumer surplus compared to single channel (see Lemmas 11, 13 & 15). However, using the Impatient scenario, we show that, when evaluating the consumer surplus for app users and non-app users combined, there is no channel arrangement that necessarily works best (see Figure 2.3 (c)). Depending on the system parameters and customer primitives, either single channel system or omnichannel systems could deliver higher combined consumer surplus than the other.

2.5. Computational Study

In this section we computationally analyze single channel and omnichannel systems with more general arrival and service processes. In particular, we assume that the number of non-app users who arrive at the store each period, and the number of service slots generated in each period follow independent Poisson distributions. Throughout the rest of this section we keep valuation, $v = 50$, and mean service rate, $\mu = 2$, fixed. We illustrate that the qualitative features of the constrained system studied in §2.4 extend to setting with a wide range of parameters of models including customer primitives, and provide additional insights. We provide the details on the computational techniques to calculate steady state in Appendix B.2.

2.5.1. Customer Strategy and Throughput

In Section 2.4.1 we observed that keeping all other parameters fixed, with increasing quality sensitivity, app users in omnichannel (with information) system become more



(a) Average Throughput for App Users

(b) Online Ordering Strategy, θ

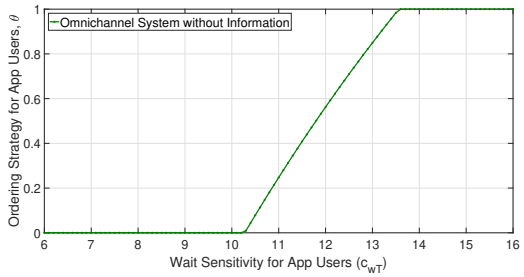
Figure 2.4. Plot (a) & (b) illustrate the non-monotonicity of average throughput and the online ordering strategy respectively, for app users in an omnichannel system without information. For illustration, the parameters are set to $c_q = 26$, $c_{wN} = 1$ and $\Lambda_N = 0.3$.

likely to choose the offline option (see Propositions 5 and 6). However, with increasing wait sensitivity, change in the app users' ordering strategy, θ, is not as straightforward. For a lower wait sensitivity, compared to quality sensitivity, app users are more likely to not order online to avoid incurring the quality penalty. Keeping all other parameters fixed, with increasing wait sensitivity, the possibility of waiting in store for order completion plays a more dominant role in app users' ordering decision. Thus, increasingly more app users order online (see Figure 2.5a) to avoid incurring waiting cost. However, if app users are highly wait-sensitive then they might not join a congested system. As a result, app users delay making an ordering decision until reaching the store and checking out how the queue has evolved. Thus, the fraction of app users ordering online eventually starts to decrease (see Figure 2.6a) with increasing wait-sensitivity. This leads to the following observation regarding the ordering strategy for app users in an omnichannel system.

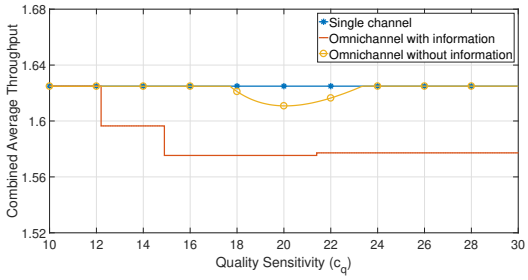
Observation 1. *Ordering strategy, θ, and hence, the average throughput for app users in an omnichannel system without information is non-monotone in app users' wait sensitivity.*

We compare this observation with the linearly decreasing demand function that is assumed in Gao and Su (2017b). As illustrated in Figure 2.4, we show that the average throughput (hence the demand), that emerges out of the equilibrium customer behavior, might increase as app users become more wait-sensitive, and exhibits a non-linear form.

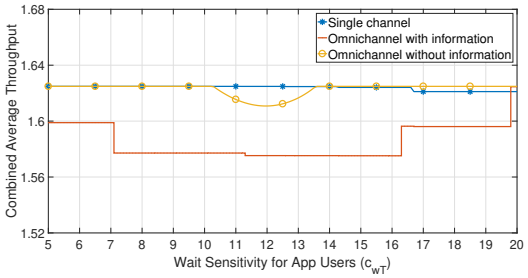
For comparison of the average throughput between a single channel and the omnichannel systems, we consider two different scenarios. In Figure 2.5, the non-app users are more



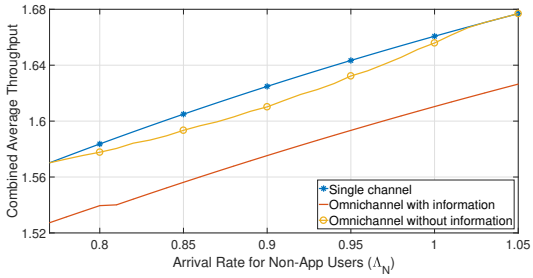
(a) App Users' Ordering Strategy, θ



(b) Combined Average Throughput (with Varying Quality Sensitivity)



(c) Combined Average Throughput (with Varying App User Wait Sensitivity)

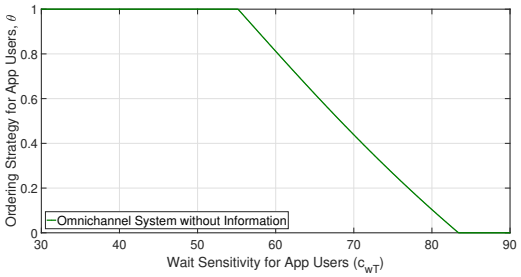


(d) Combined Average Throughput (with Varying Non-App User Arrival Rate)

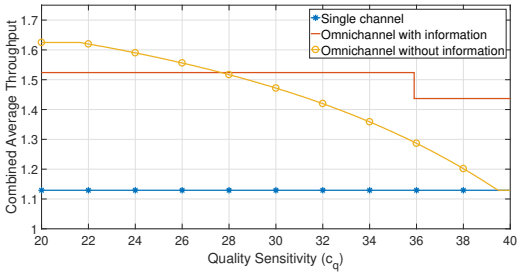
Figure 2.5. In these plots, non-app users are more wait-sensitive compared to app users, i.e. $c_{wT} < c_{wN}$. Plot (a) shows the ordering strategy, θ, for app users as a function of c_{wT}. Plots (b), (c) and (d) compare the combined (app users and non-app users) average per period throughput across single channel and omnichannel systems by varying c_q, c_{wT} and Λ_N respectively. The base value of the parameters are set to $c_q = 20$, $c_{wT} = 12$, $c_{wN} = 30$ and $\Lambda_N = 0.9$.

wait-sensitive compared to the app users, whereas in Figure 2.6, the app users are more wait-sensitive compared to the non-app users. We observe the following:

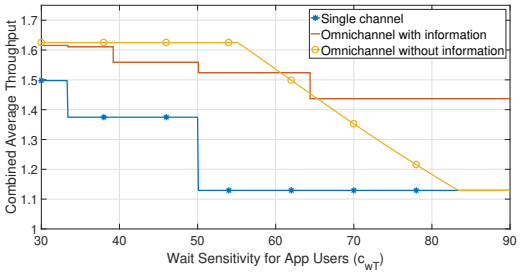
Observation 2. *(i) If non-app users are more wait-sensitive compared to app users, then single channel system may deliver higher average throughput compared to the omnichannel systems.*



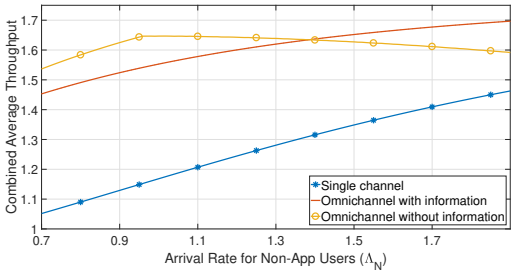
(a) App Users' Ordering Strategy, θ



(b) Combined Average Throughput (with Varying Quality Sensitivity)



(c) Combined Average Throughput (with Varying App User Wait Sensitivity)



(d) Combined Average Throughput (with Varying Non-App User Arrival Rate)

Figure 2.6. In these plots, app users are more wait-sensitive compared to non-app users, i.e. $c_{wT} > c_{wN}$. Plot (a) shows the ordering strategy, θ, for app users as a function of c_{wT}. Plots (b), (c) and (d) compare the combined (app users and non-app users) average per period throughput across single channel and omnichannel systems by varying c_q, c_{wT} and Λ_N respectively. The base value of the parameters are set to $c_q = 20$, $c_{wT} = 54$, $c_{wN} = 30$ and $\Lambda_N = 0.9$.

(ii) If app users are more wait-sensitive compared to non-app users, then omnichannel systems deliver higher average throughput compared to the single channel system.

We consistently observe that as we move from single channel to omnichannel systems, non-app users are crowded out by the app users. If non-app users are relatively more wait-sensitive, we find that enough non-app users may balk in an omnichannel system

that the overall throughput falls in comparison to a single channel system. This result is illustrated in Figure 2.5 for a range of app users' wait sensitivity, quality sensitivity and different levels of congestion in the system (we recall that, in our model, we assume only one app user arrives at the market each time period, and as a result, we alter the level of congestion in the system by altering the arrival rate for non-app users only). On the other hand, if app users are relatively more wait-sensitive, then offering online ordering option increases system throughput, as illustrated in Figure 2.6. We note that Observation 2(i) depends on the capacity of the system relative to the wait-sensitivity of app users. When there is ample capacity, having non-app users crowded out negatively affects throughput. For low capacity, the system utilization could become high enough to outweigh the effect of the loss in non-app user throughput. Thus, for low enough system capacity, the server remains busy mostly with online orders from app users. Indeed, the thresholds τ_n and τ_s could be such that neither non-app users nor app users join the queue in store. However, app users might still order online resulting in a positive throughput for omnichannel system, as compared to zero throughput for single channel. We refer to Appendix B.3 for details.

Exploring the effects of disclosing queue information on the average throughput in omnichannel systems, as illustrated in Figure 2.6, leads us to the following observation:

Observation 3. *(i) Which omnichannel information setup delivers the highest combined average throughput depends on the parameters of the system.*
(ii) Disclosing queue length information in an omnichannel system increases the combined average throughput if app users are either highly wait sensitive or highly quality sensitive or the system is highly congested.

In the absence of queue information, based on the rational belief about the steady state of the queue, app users who are highly wait-sensitive or highly quality-sensitive are deterred from ordering online. Disclosing queue information lures them into ordering online in favorable states of the queue, thereby increasing system throughput. This observation is echoed in our analytical characterization of the average throughput in Proposition 7, as illustrated in Figure 2.2a. This result also resonates with a similar message by Chen and Frank (2004) and Hassin and Roet-Green (2018). However, we extend their result from single channel to omnichannel setting with multiple customer types.

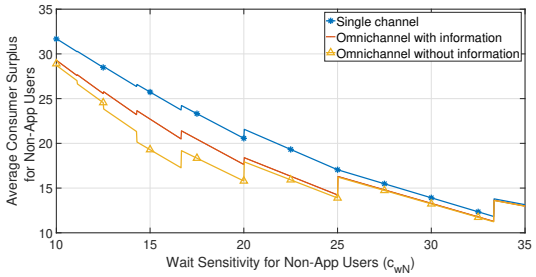
Thus, from the firm's perspective, there is no silver bullet. No channel arrangement delivers the highest throughput (and thus revenue) in all parameter regimes.

2.5.2. Consumer Surplus

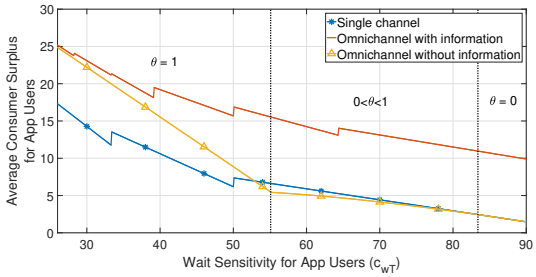
In this section, we compare the average consumer surplus for app users and non-app users across single channel and omnichannel systems, as summarized in the following observation:

Observation 4. *(i) Single channel system delivers the highest average consumer surplus for non-app users.*
(ii) Average consumer surplus for app users in a single channel system may be higher compared to an omnichannel system without information for moderate to high wait sensitivity for app users.

The presence of app users in omnichannel systems negatively impacts non-app users, and hence, single channel system dominates the omnichannel systems in terms of the average consumer surplus for non-app users. We observe this in Figure 2.7a.



(a) Average Consumer Surplus for Non-App Users



(b) Average Consumer Surplus for App Users

Figure 2.7. The above plots compare the average consumer surplus for non-app users and app users, across single channel, omnichannel (with information) and omnichannel (without information). The base values of the parameters in this figure are set to $c_q = 20$, $c_{wT} = 25$, $c_{wN} = 30$ and $\Lambda_N = 0.9$.

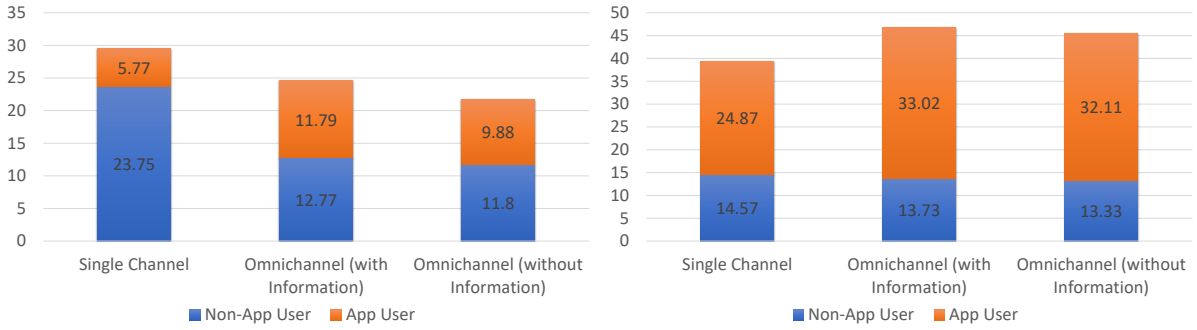
Surprisingly, we find that the app users, on the other hand, might not necessarily benefit from an omnichannel system, as illustrated in 2.7b. In the absence of queue information, app users arriving at the market makes ordering decisions based on their rational beliefs about the steady-state queue. If app users are moderately to highly wait-sensitive, ordering online can result in loss of consumer surplus since app users might join the system in unfavorable states with long queue. In contrast, app users walking in to a single channel system could balk when the queue is unfavorably long. In this sense, single channel systems offers visibility of congestion. This results in a trade-off between visibility in a single channel system and the advantage of ordering early in an omnichannel system without information (Baron et al., 2019 considers a similar trade-off). As a result, higher app user wait-sensitivity could lower the average consumer surplus for omnichannel system (without information) in comparison to single channel. Disclosing queue length information would resolve this trade-off, and result in an increase in the

average consumer surplus for app users. We illustrate this result in Figure 2.7b where we note that, for example when $c_{wT} = 55$, consumer surplus is higher in the single channel system compared to an omnichannel setting without information.

The main takeaway from this section is summarized in the following observation:

Observation 5. *Neither single channel systems nor omnichannel systems dominate one another in terms of the combined (non-app users and app users) average consumer surplus.*

In Figure 2.8b, the arrival rate for non-app users is half that of the app users' arrival rate. In this case, we observe that the single channel is dominated by the omnichannel systems in terms of consumer surplus. In contrast, as illustrated in Figure 2.8a, single channel delivers highest consumer surplus when the proportion of non-app users in the



(a) $\Lambda_N = 2$, $\Lambda_T = 1$ (b) $\Lambda_N = 0.5$, $\Lambda_T = 1$

Figure 2.8. Combined (non-app users and app users) average consumer surplus across single channel, omnichannel (with information) and omnichannel(without information). In plot (a), $\Lambda_N > \Lambda_T$, and in plot (b) $\Lambda_N < \Lambda_T$, where Λ_N and Λ_T are arrival rates for non-app users and app users respectively. For illustration purposes, the parameters are set to $c_{wN} = 20$, $c_{wT} = 22$ and $c_q = 10$.

system is higher. Hence, we need to consider the relative proportion of app users and non-app users in the system, when comparing the consumer surplus for both customer types combined. Indeed, it is possible that both customer segments are worse-off when online ordering is offered.

2.6. Conclusion

Omnichannel services are increasingly commonplace; most quick-service restaurants have integrated the physical with the digital spaces. While these novel systems aim to reduce customer wait by drawing in demand through the online ordering option, not much is known about their design implications on throughput of the system and consumer surplus generated. In particular, the effect of queue information disclosure on the system is not well understood. We adopt a discrete-time queuing model to compare single channel and omnichannel systems (with and without information provision), in terms of customer strategy, throughput, and consumer surplus. Additionally, we consider a richer customer utility model by incorporating customers' sensitivity to product quality.

We find that, the interplay between quality sensitivity and wait sensitivity results in a dual-threshold policy for app users' channel choice. In terms of delivering highest throughput, either single channel or omnichannel systems could dominate depending on the relative wait-sensitivities of app users and non-app users. Moreover, in an omnichannel system, the firm could use information disclosure as an operational lever to increase throughput when app users are either highly quality sensitive or highly wait-sensitive, or if the system has high congestion. Thus, we find that from the firm's perspective there is no silver bullet; no channel arrangement delivers the highest throughput for all

system parameters. From the customers' perspective, we once again find that neither the omnichannel nor the single channel system dominates the other in terms of the average consumer surplus for both type of customers combined. The overall consumer surplus depends on the relative proportion of app users and non-app users in the system. Indeed, it is possible that both segments are worse-off when online ordering is offered.

A common thread in our findings is that the impact of developing an online channel depends on the reaction of non-app users. Almost by definition, offering online ordering option disadvantages these consumers. While app users have the choice of when to order, non-app users are constrained to walking into the store where they may find themselves crowded out by app users. Thus, forecasting the overall performance of omnichannel systems requires a careful calibration of customer primitives and consideration of the relative proportion of non-app users in the system, in comparison to app users.

Our results hinge on the behavior of the non-app users even though several of our assumptions actually favor these customers. For example, we have used the simple first-in, first-out queue to model the service system. In reality customers may face a tandem queue, first queuing to place their order and then having their orders queued for processing. Mobile orders would be placed directly in the processing queue. A customer at the store deciding whether to enter the order queue would have to consider both all the work in front of her as well as mobile orders that might arrive to the processing queue before she gets to place her order. It is worth noting that moving to a tandem arrangement would only amplify our key result that omnichannel systems adversely impact those using the conventional channel and that the loss of these customers may outweigh gains from online sales.

A related argument could be made about relaxing our assumption that there is no cost to visiting the store (Baron et al., 2019 and Roet-Green and Yuan, 2019 make similar assumptions) in person. Because of this assumption app users in our model do not balk until they have traveled to the physical store and observed the queue. That is, they have a choice of when to order. That still holds if there is non-zero traveling cost. However, it could complicate the app users' ordering strategy, for example, app users could potentially randomize between ordering online, choosing the offline option, and balking (before going to the store). In this sense, a zero travel cost allows parsimony without a significant loss of insight on app user behavior. Moreover, a non-zero traveling cost for all customers would adversely affect non-app users. They would now have to evaluate whether the expected net benefit of going to the store outweighs the cost to get there. A channel structure that increases the number of app users in the system could then result in fewer non-app users visiting the store. In our current model, the number of app users visiting the store is independent of the channel structure; alternative structures just result in different numbers of non-app users joining the queue. Positive travel costs could exacerbate this by having fewer non-app users actually enter the store.

In our paper, we consider two type-dependent wait sensitivities. Future research could explore the effect of incorporating a higher degree of heterogeneity within each customer type. With respect to information disclosure, consistent with the literature, we focus on queue length information. One could extend the analysis to explore the effect of disclosing other type of congestion information, for e.g. expected wait times, to the app users. Finally, using our model to explore the effect of service prioritization in the presence of quality sensitive customers could be a worthwhile direction for future research.

CHAPTER 3

To App or Not to App: Omnichannel Competition Among Retail Services (Joint work with Achal Bassamboo and Martin Lariviere)

3.1. Introduction

Large firms in the quick-service industry have taken the lead in developing their own digital ordering app. Major players such as Starbucks and Dunkin' have been among the early adopters of this technology. But even to this day, other players in the industry are investing resources to develop digital infrastructure in order to offer digital apps to their customers. While having the app-ordering option is appealing to customers, who seem to value the convenience of ordering and payment it provides, still not all players in the industry offer an app. It is not clear whether it is the fixed cost of developing the digital infrastructure that is deterring these firms from offering apps, or is there something more? We examine whether offering apps to customers necessarily always leads to higher revenue for firms in a competitive market even if we ignore the upfront cost of implementing an app.

In a mature competitive market, through repeated interactions customers develop loyalty towards particular firms (see Klein, 2018). Firm loyalty could arise from a variety of factors such as: customers' affinity towards specific product attributes, convenience arising from location of a store or due to any number reasons pertaining to customer

idiosyncrasies. For example, while customer A might prefer Starbucks over a locally owned independent coffee shop (say, Chicago-based Backlot Coffee), because the Starbucks store is on the way of her daily commute, customer B might prefer Backlot Coffee because she likes a particular blend of coffee that is available only at Backlot. In this setting, each customer when they are in the market for coffee, prefers being served by their preferred store over the other. However, since most customers are delay-sensitive, they might be willing to receive service from their second-choice if their first-choice is highly congested. In this sense, the firms serve a *segmented market* where each customer in the market has a preference to be served by a particular firm and visit the other alternative if the waits at the preferred firm is too long.

In a segmented and competitive market, the level of congestion could depend on whether or not firms offer apps, since availability of apps affects customers' ordering strategy. In particular, there are two aspects of app-ordering that might affect congestion: *advanced-ordering effect* and *information disclosure*. Advanced-ordering effect simply refers to customers' ability to order whenever and from wherever they want. Information disclosure refers to the practice of communicating congestion information to customers via the app. This assumption is motivated by firms such as Starbucks which offer estimated waiting time to customers via the app before they decide whether or not to order. While Backlot stores are often located near Starbucks outlets, Backlot Coffee does not offer app even though Starbucks does. Now, consider a loyal Starbucks customer opening the app on her phone and realizing long waiting-time. Should this customer still place an order for her favorite coffee with Starbucks, possibly ending up waiting a long time before getting her coffee? Or should she consider walking-in to her second choice, i.e. Backlot Coffee, on

the off-chance that it is less congested compared to Starbucks? This trade-off is central to the analysis in our work.

We present a game-theoretic framework to analyze and compare the revenue for two retail firms when they compete with each other on whether to offer a digital ordering app or not (even if we ignore the upfront cost of implementing an app). In this work we address the following research question: *Does offering a digital ordering app always leads to an increase in firms' revenue in a competitive setting?*

3.2. Literature Review

There is an extensive literature on markets where servers compete over delay-sensitive customers by posting prices. Most of these papers use a two-stage game wherein the first stage servers announce prices, following which customers in the second stage select servers accordingly (see Hassin and Haviv, 2003). Our work differs from this literature in that we consider a market where the firms are price-takers. Instead of price competition, we consider a competitive setting where we examine to what extent firms' revenue changes upon offering digital ordering apps, which provide congestion information and advanced ordering options.

Our work is also related to the literature on competition in availability. Dana Jr and Petruzzi (2001), considers a setting where customers incur a cost in case of a stock-out, which affects firm's inventory decisions. Lippman and McCardle (1997), and Mahajan and Van Ryzin (2001) consider similar markets where customers exogenously decide which firm to visit first but they may switch among competitors if the first firm stocks-out. In our work we consider a mature market so that customers have a preference of being served by

a particular firm. However, a congested system might drive customers to the competitor. Moreover, our paper considers firm's capacity as exogenously given and focuses on whether or not to offer digital apps, which affect customer ordering decision through availablity of congestion information.

There is body of research in revenue management that considers competitive settings in airline industry, focusing on implications of competitive allocation on seat inventory control (see Netessine and Shumsky, 2005). In these settings, the allocation of seat inventory among fare classes by one airline affects the quantity of customer demand and optimal seat allocation of the other airline. In Netessine and Shumsky (2005), the decision is to allocate fixed capacity among two customer classes. In contrast, in our work instead of explicit allocation rules, the availibility of apps determine the extent to which capacity gets allocated among app-users and walk-in customers.

Finally, our work is closely related to the work on restaurant reservation by Alexandrov and Lariviere (2012). They address the question of whether or not restaurants should offer reservations. They extend their analysis to a competitive environment. The focus in their work is on reservations' ability to influence customer behavior and thus increase sales. Similar to our work, Alexandrov and Lariviere (2012) considers market uncertainty with two demand states. In our work, a high demand realization might affect customers' willingness to walk-in to the firm due to anticipated congestion. This can be counteracted by offering apps where customers could order in advance. However, in a competitive setting it is not clear whether offering an app necessarily leads to higher revenue for a firm.

3.3. Model

In this section we present a game-theoretic model for analyzing competition among firms for offering digital apps to customers.

Market. The market is comprised of two revenue-maximizing firms indexed by $i = 1, 2$. We denote the competitor to firm i by i^-. The capacity of firm i to serve customers is denoted by $\mu_i > 0$ where $i = 1, 2$. Every customer in the market has an a priori preference to be served by one of the two firms. If a customer is served by her most preferred firm, her valuation for service is denoted by $v > 0$. On the other hand, if she is served by her second-choice, then her valuation for service is denoted by $\hat{v}$, where $0 < \hat{v} < v$. In this sense, the market is *segmented* into *loyal* and *non-loyal* customers. The size of the customer segment comprised of customers loyal to firm i is denoted by Λ_i where $i = 1, 2$. All customers in a given segment are homogeneous. The realization of the market sizes is random, i.e. Λ_i is a random variable which takes values H_i and L_i with probability p_i and $1 - p_i$ respectively, where $0 < L_i < H_i$. We assume that the realizations for Λ_1 and Λ_2 are independent of each other. The parameters p_i, L_i and H_i are common knowledge to all the customers and the two firms in the market. Finally, all customers in the market are delay-sensitive. We model customers' waiting cost such that customers loyal (or non-loyal) to firm i seek service with firm i as long as their position in line (if they were to seek service), among all the customers seeking service, is not greater than a threshold, τ_i (or $\hat{\tau}_i$). In particular, let X_i denote the line position of a customer if she were to join the queue for firm i. We model waiting cost $c_i(X_i)$ (or $\hat{c}_i(X_i)$) for a customer loyal (or

non-loyal) to firm i as follows:

$$c_i(X_i) = \begin{cases} 0 & if\ X_i \leq \tau_i \\ & \\ \infty & if\ X_i > \tau_i \end{cases}, \qquad \hat{c}_i(X_i) = \begin{cases} 0 & if\ X_i \leq \hat{\tau}_i \\ & \\ \infty & if\ X_i > \hat{\tau}_i \end{cases}.$$

We assume that $\hat{\tau}_i < \tau_i$ for $i = 1, 2$. We can think of τ_i and $\hat{\tau}_i$ to be implicitly linked to either the capacity of the firm or customers' valuation for service; a customer might be willing to join a longer line if the firm has larger capacity, or customers have a higher valuation for service. An exact relationship of τ_i or $\hat{\tau}_i$ with capacity or valuation is a modeling choice. In our model we do not explicitly assume any such relationship (except in §3.4 Figure 3.2). Furthermore, we focus our analysis on the case where the thresholds satisfy the following condition:

$$L_i + 1 \leq \tau_i < H_i \tag{3.1}$$

for $i = 1, 2$. Condition (3.1) captures the setting where a high demand realization corresponds to a congested system, such that not all loyal customers choose to join service. On the other hand, a low demand realization corresponds to a lightly loaded system. Before the game begins, each of the two firms decides whether or not to offer an app. This decision is common knowledge to all the customers and both firms in the market.

Game. Now, we describe the game. We propose a two-period game.

Period 1. First, the market sizes, Λ_1 and Λ_2, are realized but not necessarily observed. Depending on whether or not firm i offers app, there are two scenarios:

(i) *App.* If firm i offers an app, all customers loyal to firm i open the app. These customers simultaneously decide whether or not to order through the app. We can think about the app-ordering process as follows. First, these customers are uniformly randomly ordered. Then, depending upon the threshold τ_i and the demand Λ_i, customers up to τ_i order through the app. Along with condition (3.1), this implies that for a low demand realization of $\Lambda_i = L_i$, all L_i loyal customers seek service with firm i through the app. However, for a high demand realization of $\Lambda_i = H_i$, only τ_i out of H_i loyal customers join the queue for service through the app. Each of the remaining $H_i - \tau_i$ customers randomly decide between firm i^- and balking at the start of Period 2. We model this decision by each customer independently flipping a Bernoulli coin with probability θ_i, where θ_i represents the probability that the customer walks-in to firm i^- in Period 2. Probability that the customer balks at the outset is $1-\theta_i$. Henceforth, we use the notation $B(n, p)$ to denote a Binomial random variable with n trials and success probability p. The total number of customers who walk-in to firm i^- in Period 2 is distributed according to a Binomial random variable, $B(n, p)$ where $n = H_i - \tau_i$ and $p = \theta_i$.

By opening the app, all customers loyal to firm i gain perfect knowledge of the realization of Λ_i. However, this knowledge does not resolve their uncertainty regarding the realization of the market size Λ_{i^-}, the customer segment loyal to competitor firm i^-.

(ii) *No App.* If firm i does not offer app, customers loyal to firm i remain uncertain about the realizations of both customer segment sizes Λ_i and Λ_{i^-}. In the absence

of an app, customers loyal to firm i cannot order in Period 1. They are faced with two decisions: (i) first they randomly decide whether to balk in Period 1 and not patronize either of the two firms in Period 2, or (ii) to not balk in Period 1 and subsequently randomly decide which one of the two firms to visit in Period 2. We use α_i to denote the probability with which each of the Λ_i customers does not balk in Period 1. We further use θ_i to denote the probability with which each of the customers, who does not balk, visits the competitor firm i^- in Period 2. Thus, for each customer there are three possible random outcomes: (i) balk in Period 1 with probabilily $1-\alpha_i$, (ii) do not balk in Period 1 and visit firm i^- in Period 2 with probability $\alpha_i\theta_i$, and (iii) do not balk in Period 1 and visit firm i in Period 2 with probability $\alpha_i(1-\theta_i)$. We will henceforth use the notation $O(\Lambda_i, 1-\alpha_i)$ to denote the random number of customers out of Λ_i customers who balk in Period 1, we will use $O(\Lambda_i, \alpha_i\theta_i)$ to denote the random number of customers who do not balk in Period 1 and visit firm i^- in Period 2, and finally we will use $O(\Lambda_i, \alpha_i(1-\theta_i))$ to denote the random number of customers who do not balk in Period 1 and visit firm i in Period 2, where the random vector $(O(\Lambda_i, 1-\alpha_i), O(\Lambda_i, \alpha_i\theta_i), O(\Lambda_i, \alpha_i(1-\theta_i)))$ follows a Multinomial distribution with Λ_i trials and the vector of success probabilities $(1-\alpha_i, \alpha_i\theta_i, \alpha_i(1-\theta_i))$.

Period 2. Depending upon whether or not each of the two firms offer apps, some fraction of customers in Period 1 visit the two firms in Period 2 according to their decisions in Period 1, as described above. As a result, first, the markets in Period 2 are realized for each of the two firms. Each of these markets are potentially comprised of loyal as well as non-loyal customers corresponding to that firm. Subsequently, all customers in each

of the Period 2 markets, simultaneously decide whether to seek service or to balk. All services take place at the end of Period 2. For any given firm i, if there are customers who joined the queue in Period 1, they are served before the service for Period 2 customers takes place.

Before we present customer utilities, let us describe the general problem of customers joining service in Period 2. We will use the notation in this description to represent the customer utilities, and use the solution for this problem to solve the two-period game.

Description of Period 2 ordering problem. For a given firm i, let M_i and $\hat{M}_i$ denote the number of loyal and non-loyal customers, respectively, who visit firm i in Period 2 (walk-in customers). Also, let q_i denote the initial number of customers who joined the queue at firm i in Period 1. All customers in the Period 2 market are ordered uniformly randomly before they decide whether to join firm i or balk. Given any ordering, all non-loyal customers seek service up to a threshold $\hat{\tau}_i$. Similarly, all loyal customers seek service up to threshold τ_i. All remaining customers balk. Given thresholds τ_i and $\hat{\tau}_i$ and the realized markets, the probability that a customer loyal to firm i joins service in Period 2 with firm i, is denoted by,

$$\pi^{\tau_i,\hat{\tau}_i}(q_i, M_i, \hat{M}_i). \tag{3.2}$$

Similarly, the probability that a customer not loyal to firm i joins it's queue in Period 2 is denoted by,

$$\hat{\pi}^{\tau_i,\hat{\tau}_i}(q_i, M_i, \hat{M}_i). \tag{3.3}$$

For brevity of notation, we will henceforth use π_i interchangeably with (3.2), and $\hat{\pi}_i$ interchangeably with (3.3).

Let us consider an example for the ordering process with firm i in Period 2. Let $M_i = 4$, $\hat{M}_i = 3$, $q_i = 1$, $\hat{\tau}_i = 3$ and $\tau_i = 6$. Thus, there are four loyal and three non-loyal customers in the Period 2 market for firm i, and there is one customer who joined service with firm i in Period 1. Let $n_1 n_2 l_1 n_3 l_2 l_3 l_4$ denote a possible ordering for the Period 2 customers, where l_k and n_k are the k^{th} loyal and k^{th} non-loyal customers respectively. Since we assume that Period 1 customers get served before Period 2 customers, the effective positions of customers n_1, n_2 and l_1 are 2, 3 and 4 respectively. These customers join service. Customer n_3 balks since her position is 5 and $\hat{\tau}_i = 3$. Now, since n_3 balks, effective positions for l_2, l_3 and l_4 become 5, 6 and 7. Since, $\tau_i = 6$, customers l_2 and l_3 join service, but customer l_4 balk.

Customer Utility. As already mentioned, customers' valuation for being served by their most-preferred firm and second-preferred firm are $v > 0$ and $\hat{v} > 0$ respectively. In addition, here we introduce cost $D > 0$ that the customers incur when they visit a firm in Period 2 but do not seek service due to congestion. This represents the hassle cost of being turned away. All customers who balk in Period 1 get zero utility.

First, let us focus on utilities in Period 2 for customers who visit either of the two firms in Period 2. As per our description of Period 2 above, these customers are first uniformly randomly ordered, and then they either join service or balk depending on their position in the ordering. Thus, the utility of a loyal firm i customer who ends up in the

market for firm i in Period 2 is given by,

$$u_i(q_i, M_i, \hat{M}_i) = \pi_i \cdot v + (1 - \pi_i) \cdot (-D). \tag{3.4}$$

where, as per (3.2), π_i represents the probability that in Period 2 a loyal firm i customer joins service with firm i. Similarly, the utility of a customer who is not loyal to firm i and ends up in the market for firm i in Period 2 is given by,

$$\hat{u}_i(q_i, M_i, \hat{M}_i) = \hat{\pi}_i \cdot \hat{v} + (1 - \hat{\pi}_i) \cdot (-D). \tag{3.5}$$

We note that (3.4) and (3.5) depend on the realizations of q_i, M_i and $\hat{M}_i$ through π_i and $\hat{\pi}_i$. Now, depending on whether or not each of the two firms offer an app, there are four scenarios which is denoted by (y_i, y_{i^-}) where $y_i, y_{i^-} \in \{\mathrm{A}, \mathrm{NA}\}$ are the decisions of firm i and firm i^- respectively (A stands for app and NA stands for no app). For a customer in Period 1 who is loyal to firm i, we present the expected utility associated with all possible actions in each of the four scenarios:

Scenario (A,A): As already mentioned, if $\Lambda_i = L_i$, all of these L_i loyal customers join firm i through the app in Period 1. If $\Lambda_i = H_i$, there are $B(H_i - \tau_i, \theta_i)$ loyal customers who visit firm i^- in Period 2, where $B(n, p)$ denotes a Binomial random variable with n trials and success probability p. From (3.5), utility associated with this action is $\hat{u}_{i^-}(q_{i^-}, M_{i^-}, \hat{M}_{i^-})$, where $M_{i^-} = 0$, $\hat{M}_{i^-} = B(H_i - \tau_i, \theta_i)$, and $q_{i^-} = L_{i^-}$ if $\Lambda_{i^-} = L_{i^-}$ and $q_{i^-} = \tau_{i^-}$ if $\Lambda_{i^-} = H_{i^-}$. Thus, the expected utility is given by,

$$\mathbb{E}_{B,\Lambda_{i^-}}[\hat{u}_{i^-}(\min(\tau_{i^-}, \Lambda_{i^-}), 0, B(H_i - \tau_i, \theta_i))]. \tag{3.6}$$

Scenario (A,NA): Similar to Scenario (A,A), if $\Lambda_i = L_i$, all L_i loyal customers join firm i through the app in Period 1. If $\Lambda_i = H_i$, then $B(H_i - \tau_i, \theta_i)$ loyal firm i customers visit firm i^- in Period 2, and the remaining customers balk. Since, firm i^- does not offer an app in this scenario, $M_{i^-} = O(\Lambda_{i^-}, \alpha_{i^-}(1-\theta_{i^-}))$ customers loyal to firm i^- visit firm i^- in Period 2. The expected utility associated with the action of a firm i loyal customer visiting firm i^- in Period 2 is given by,

$$\mathbb{E}_{B,O,\Lambda_{i^-}}[\hat{u}_{i^-}(0, O(\Lambda_{i^-}, \alpha_{i^-}(1-\theta_{i^-})), B(H_i - \tau_i, \theta_i))]. \tag{3.7}$$

Scenario (NA,A): In this scenario, since firm i does not offer app, customers loyal to firm i choose to visit either firm i or firm i^- in Period 2. We present the expected utilities associated with these two actions. Using (3.4), the expected utility for a loyal firm i customer visiting firm i in Period 2 is given by,

$$\mathbb{E}_{B,O,\Lambda_i,\Lambda_{i^-}}[u_i(0, O(\Lambda_i, \alpha_i(1-\theta_i)), B(\max(0, \Lambda_{i^-} - \tau_{i^-}), \theta_{i^-}))]. \tag{3.8}$$

Using (3.5), the expected utility for a loyal firm i customer visiting firm i^- is given by,

$$\mathbb{E}_{O,\Lambda_i,\Lambda_{i^-}}[\hat{u}_{i^-}(\min(\tau_{i^-}, \Lambda_{i^-}), 0, O(\Lambda_i, \alpha_i\theta_i))]. \tag{3.9}$$

Scenario (NA,NA): In this scenario neither firm offers an app. The expected utility of a customer loyal to firm i associated with the action of visiting firm i in Period 2 is given by,

$$\mathbb{E}_{O,\Lambda_i,\Lambda_{i^-}}[u_i(0, O(\Lambda_i, \alpha_i(1-\theta_i)), O(\Lambda_{i^-}, \alpha_{i^-}\theta_{i^-}))]. \tag{3.10}$$

Finally, the expected utility of a customer loyal to firm i associated with the action of visiting competitor firm i^- is given by,

$$\mathbb{E}_{O,\Lambda_i,\Lambda_{i^-}}[\hat{u}_{i^-}(0, O(\Lambda_{i^-}, \alpha_{i^-}(1-\theta_{i^-})), O(\Lambda_i, \alpha_i\theta_i))]. \tag{3.11}$$

In each of the four scenarios described above, customers in Period 1 choose actions that maximize their expected utility in Period 2 given by expressions (3.6), (3.7), (3.8), (3.9), (3.10) and (3.11). In all of the four Scenarios, we denote the optimal customer decisions by $\alpha_i^*, \alpha_{i^-}^*, \theta_i^*$ and $\theta_{i^-}^*$.

Throughput. Next, we present the expressions for throughput for firm i in all the four scenarios.

Scenario (A,A): In this scenario, since firm i offers an app, $q_i = \min(\tau_i, \Lambda_i)$ customers join using the app in Period 1. In Period 2, there are $M_i = 0$ customers from firm i and $\hat{M}_i = B(\max(0, \Lambda_{i^-} - \tau_{i^-}), \theta_{i^-})$ customers from firm i^-, where $B(n,p)$ denotes a Binomial random variable with n trials and success probability p. Thus, firm i's throughput is given by,

$$\mathbb{E}_{B,\Lambda_i,\Lambda_{i^-}}[q_i + \hat{\pi}_i(q_i, 0, \hat{M}_i) \cdot \hat{M}_i]. \tag{3.12}$$

Scenario (A,NA): Firm i's throughput in this scenario is given by,

$$\mathbb{E}_{O,\Lambda_i,\Lambda_{i^-}}[q_i + \hat{\pi}_i(q_i, 0, \hat{M}_i) \cdot \hat{M}_i]. \tag{3.13}$$

where $q_i = \min(\tau_i, \Lambda_i)$, $M_i = 0$ and $\hat{M}_i = O(\Lambda_{i^-}, \alpha_{i^-}\theta_{i^-})$.

SCENARIO (NA,A): In this scenario, since firm i does not offer an app, no customers join firm i in Period 1. Thus, $q_i = 0$. In Period 2, there are $M_i = O(\Lambda_i, \alpha_i(1-\theta_i))$ customers from firm i and $\hat{M}_i = B(\max(0, \Lambda_{i^-} - \tau_{i^-}), \theta_{i^-})$ customers from firm i^-. Hence, the throughput for firm i is given by,

$$\mathbb{E}_{B,O,\Lambda_i,\Lambda_{i^-}}[\pi_i(0, M_i, \hat{M}_i) \cdot M_i + \hat{\pi}_i(0, M_i, \hat{M}_i) \cdot \hat{M}_i]. \tag{3.14}$$

SCENARIO(NA,NA): Finally, the throughput for firm i in this scenario where none of the two firms offer an app, is given by,

$$\mathbb{E}_{O,\Lambda_i,\Lambda_{i^-}}[\pi_i(0, M_i, \hat{M}_i) \cdot M_i + \hat{\pi}_i(0, M_i, \hat{M}_i) \cdot \hat{M}_i]. \tag{3.15}$$

where $M_i = O(\Lambda_i, \alpha_i(1-\theta_i))$ and $\hat{M}_i = O(\Lambda_{i^-}, \alpha_{i^-}\theta_{i^-})$.

In the following section, we numerically compute the optimal customer decisions, $\alpha_i^*, \alpha_{i^-}^*, \theta_i^*$ and $\theta_{i^-}^*$, for specific scenarios and parameter values. We then use it to numerically compute the throughput under the scenarios considered, in order to evaluate the firm's optimal decision.

3.4. Numerical Results, Discussion and Future Work

In this section, we examine under what parameter regime offering an app increases throughput for a firm when its competitor offers an app. We consider throughput as a proxy for the firm's revenue. We compare the throughput for firm 2 under Scenarios (A,A) and (NA,A) where firm 1 always offers an app. In particular, for simplification of our analysis, we examine these two scenarios by setting the value of the parameter $D = 0$, which is the hassle cost of visiting the store in Period 2 but not being served. Thus, no

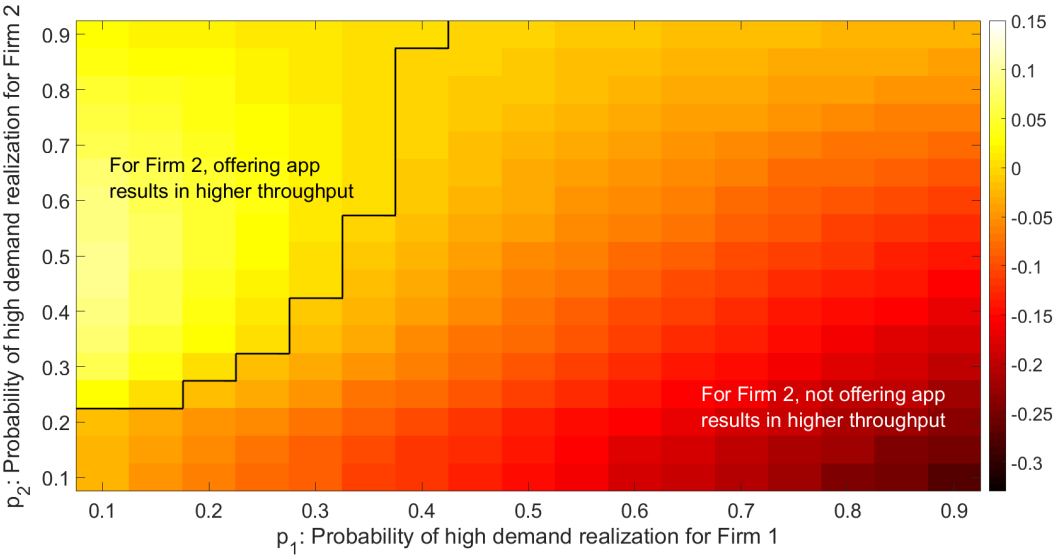


Figure 3.1. Value of the difference in throughput for Firm 2 between the two Scenarios (A,A) and (A,NA). Positive values in the color scale correspond to higher throughput when Firm 2 offers an app, i.e. Scenario (A,A). Negative values correspond to higher throughput when Firm 2 does not offer an app, i.e. Scenario (A,NA). Parameter values: $v = 1$, $\hat{v} = 0.9$, $\tau_1 = \tau_2 = 5$, $\hat{\tau}_1 = \hat{\tau}_2 = 4$, $L_1 = L_2 = 1$, $H_1 = H_2 = 10$ and $D = 0$.

customer in Period 1 balks at the outset, and always optimally decides to visit one of the two firms in Period 2. This implies $\alpha_2^* = 1$ and $\theta_1^* = 1$, when firm 1 always offers an app in Scenarios (A,A) and (NA,A). We numerically compute θ_2^* using expressions (3.8) and (3.9), and use it to compute firm 2's throughput using expressions (3.12) and (3.14) under optimal customer decisions. Figure 3.1 and Figure 3.2 specify the values of the parameters of the model that we solve for, and illustrate our findings.

In Figure 3.1, we look at the difference in firm 2's throughput between Scenario (A,A) and Scenario (NA,A), where firm 1 always offers an app. The X-axis and the Y-axis represent the probabilities, p_1 and p_2, with which firm 1 and firm 2 respectively have a high demand realization. The region in Figure 3.1 with positive values according to

the color-scale correspond to parameters values (p_1, p_2) where offering an app leads to higher throughput for firm 2. We keep all other parameters same for both firms. We observe that, when it is better for firm 2 to offer an app, the probability of high demand realization is relatively lower for firm 1 compared to firm 2 when all other parameters of the model are same for both firms. In the absence of an app, firm 2 customers do not observe the realized demand for firm 2 in Period 1. As a result, in expectation firm 2 customers are more likely to visit firm 1 in Period 2 because they anticipate a higher likelihood of being served. Offering an app puts firm 2 customers at an advantage; they get to see the demand realization in Period 1 and order in advance. Thus, by offering an app, firm 2 keeps loyal customers from defecting. In this case, firm 2's decision to offer an app is driven by retaining its loyal customers, given that there are not enough firm 1 customers who would potentially visit firm 2 in Period 2 due to a low probability of a high market realization for firm 1.

However, when the probability of high demand realization for firm 1 is relatively high compared to firm 2, offering an app might hurt firm 2. When an app is available, some of the firm 2 customers order in advance through the app in Period 1. This could prevent potential firm 1 customers from seeking service at firm 2 in Period 2. This is true especially since firm 1 loyalists have a lower joining threshold, $\hat{\tau}_2$, compared to the threshold, τ_2, for loyal firm 2 customers. As a result, firm 1 customers are deterred from visiting firm 2 in Period 2. This results in lower throughput for firm 2 since there are relatively more firm 1 customers in the market in expectation compared to firm 2 customers. When an app is not offered by firm 2, all Period 2 customers in the market simultaneously make joining

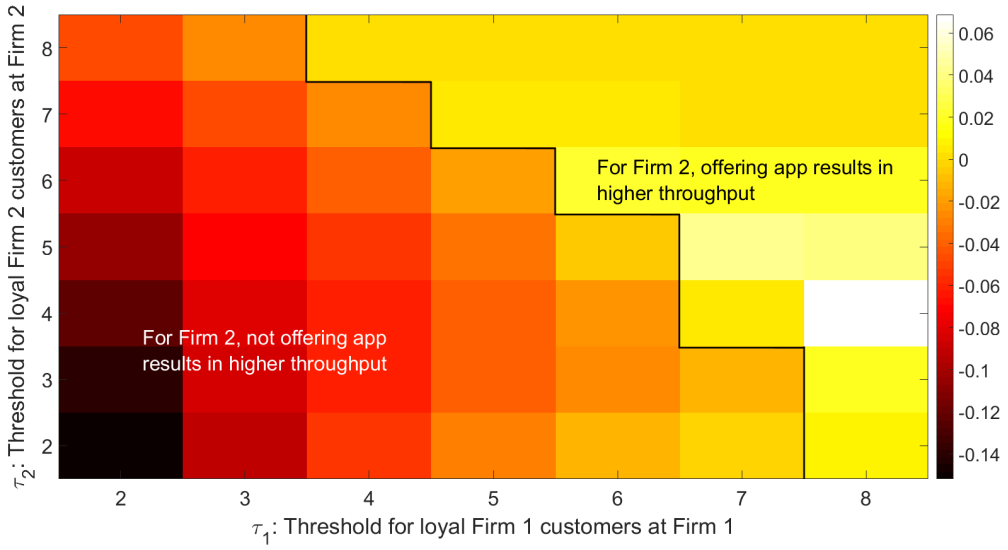


Figure 3.2. Value of the difference in throughput for Firm 2 between the two Scenarios (A,A) and (A,NA). Positive values in the color scale correspond to higher throughput when Firm 2 offers an app, i.e. Scenario (A,A). Negative values correspond to higher throughput when Firm 2 does not offer an app, i.e. Scenario (A,NA). Parameter values: $v = 1$, $\hat{v} = 0.9$, $p_1 = p_2 = 0.3$, $L_1 = L_2 = 1$, $H_1 = H_2 = 10$, $D = 0$. We assume that $\hat{\tau}_1 = \lfloor (\frac{\hat{v}}{v})\tau_1 \rfloor$ and $\hat{\tau}_2 = \lfloor (\frac{\hat{v}}{v})\tau_2 \rfloor$.

decisions, which increases the likelihood of firm 1 customers joining service with firm 2. Thus, not offering app in this case results in higher throughput for firm 2.

Again, in Figure 3.2 we look at the difference in firm 2's throughput between Scenario (A,A) and Scenario (NA,A), where firm 1 always offers an app. In this case, the X-axis and the Y-axis represent the joining thresholds for loyal customers of firm 1 and firm 2 respectively. We assume that the threshold for non-loyal customers, $\hat{\tau}_i$, weakly increases with τ_i. We can implicitly link the change in τ_i with a change in firm i's capacity. Thus, a higher value of τ_i corresponds to a firm with larger capacity. All other parameters are fixed at the same values for both firms. An insight similar to Figure 3.1 is generated

from Figure 3.2. We observe that, for a fixed value of τ_2, it is better for firm 2 to offer an app beyond a certain value of τ_1. As we increase τ_1 keeping τ_2 fixed, $\hat{\tau}_1$ increases as well. Consequently in the event of a low demand realization for firm 1, the likelihood that a loyal customer from firm 2 visiting firm 1 in Period 2 is served increases as well. Additionally, in the event of high market realization for firm 1, an increase in τ_1 results in lesser number of firm 1 customers visiting firm 2 in Period 2. Thus, offering an app allows firm 2 to retain its loyal customers and thus resulting in a higher throughput.

To summarize our work, we consider competing service providers, with the goal of examining whether offering an app always leads to an increase in the firm's revenue when the competitor firm offers an app. Availability of an ordering app affects customers' choice of firm i) by offering customers visibility to the level of system congestion, and ii) by offering customers the option to order in advance before arriving at the store. If an app is not available, customers make decisions a priori without any available information about the congestion. In this case, customers first decide whether or not to patronize either of the two firms, and if they do, they additionally decide whether to visit the preferred firm or the competitor firm. As a result, a firm could potentially lose its loyal customers to its competitor. Informing customers about congestion in the system via an app may alleviate this problem. On the other hand, if an app is available, the advanced-ordering effect may result in a higher congestion in the system, which may deter the competitor's customers from joining the system. Thus, taking into consideration the relative sizes of their customer bases, and the joining thresholds for the two customer segments, offering an app might or might not be in the firm's best interest, even if we ignore the upfront cost of implementing an app. Firms' app-offering strategy would depend, among other

factors, on the extent to which they are trying to retain their loyal customers, and the extent to which they are competing for their competitor's customers.

In our future work, using the framework described in this chapter, we would like to solve a simultaneous-move game where both firms decide whether or not to offer an app. We will analyze the equilibrium outcome of that game under various parameter regime. Our findings from the example in this section establish that the scenario where both firms offer an app might not necessarily be an equilibrium outcome of the game, and a firm's best response might be to not offer an app given the competitor's decision to offer an app. We would like to investigate whether or not there always exists an unique equilibrium. Additionally, it would be interesting to find out whether or not both firms deciding to not offer an app could possibly be an equilibrium outcome of the game.

References

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

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

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

Ahmadi-Javid A, Jalali Z, Klassen KJ (2017) Outpatient appointment systems in healthcare: A review of optimization studies. *European Journal of Operational Research* 258(1):3–34.

Alexandrov A, Lariviere MA (2012) Are reservations recommended? *Manufacturing & Service Operations Management* 14(2):218–230.

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

Ata B, Olsen TL (2013) Congestion-based leadtime quotation and pricing for revenue maximization with heterogeneous customers. *Queueing Systems* 73(1):35–78.

Baron O, Chen X, Li Y (2019) The paradox of choice: The false premise of omnichannel services and how to realize it. *Working Paper* .

Çelik S, Maglaras C (2008) Dynamic pricing and lead-time quotation for a multiclass make-to-order queue. *Management Science* 54(6):1132–1146.

Chen H, Frank M (2004) Monopoly pricing when customers queue. *IIE Transactions* 36(6):569–581.

Chen M, Hu M, Wang J (2019) Food delivery service and restaurant: Friend or foe? *Available at SSRN* .

Cil EB, Lariviere MA (2013) Saving seats for strategic customers. *Operations Research* 61(6):1321–1332.

Dana Jr JD, Petruzzi NC (2001) Note: The newsvendor model with endogenous demand. *Management Science* 47(11):1488–1497.

Doroudi S, Akan M, Harchol-Balter M, Karp J, Borgs C, Chayes JT (2013) Priority pricing in queues with a continuous distribution of customer valuations. Technical report, Technical report CM-CS-13-109, Computer Science Department, Carnegie Mellon ....

Edelson NM, Hilderbrand DK (1975) Congestion tolls for poisson queuing processes. *Econometrica: Journal of the Econometric Society* 81–92.

Feldman P, Frazelle AE, Swinney R (2019) Can delivery platforms benefit restaurants? *Available at SSRN* .

Gallino S, Moreno A (2014) Integration of online and offline channels in retail: The impact of sharing reliable inventory availability information. *Management Science* 60(6):1434–1451.

Gao F, Su X (2017a) Omnichannel retail operations with buy-online-and-pick-up-in-store. *Management Science* 63(8):2478–2492.

Gao F, Su X (2017b) Omnichannel service operations with online and offline self-order technologies. *Management Science* 64(8):3595–3608.

Guo P, Zipkin P (2007) Analysis and comparison of queues with different levels of delay information. *Management Science* 53(6):962–970.

Gurvich I, Lariviere MA, Ozkan C (2018) Coverage, coarseness, and classification: Determinants of social efficiency in priority queues. *Management Science* 65(3):1061–1075.

Hassin R (1986) Consumer information in markets with random product quality: The case of queues and balking. *Econometrica: Journal of the Econometric Society* 1185–1195.

Hassin R (2016) *Rational queueing* (CRC press).

Hassin R, Haviv M (2003) *To queue or not to queue: Equilibrium behavior in queueing systems*, volume 59 (Springer Science & Business Media).

Hassin R, Roet-Green R (2017) The impact of inspection cost on equilibrium, revenue, and social welfare in a single-server queue. *Operations Research* 65(3):804–820.

Hassin R, Roet-Green R (2018) The armchair decision: On queue-length information when customers travel to a queue. *Available at SSRN* .

Hsu VN, Xu SH, Jukic B (2009) Optimal scheduling and incentive compatible pricing for a service system with quality of service guarantees. *Manufacturing & Service Operations Management* 11(3):375–396.

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

Ibrahim R (2018) Sharing delay information in service systems: a literature survey. *Queueing Systems* 89(1-2):49–79.

Kang K, Doroudi S, Delasay M (2020) Designing efficient omni-channel services. *Available at SSRN* .

Katta AK, Sethuraman J (2005) Pricing strategies and service differentiation in queues—a profit maximization perspective. *Technical Report* .

Kelso A (2019) Chipotle doubling down on digital is paying off wildly. URL https://www.forbes.com/sites/aliciakelso/2019/10/23/how-chipotle-leveraged-its-digital-ecosystem-to-exceed-q3-expectations/.

Klein D (2018) Which chains have the most loyal customers? URL https://www.qsrmagazine.com/fast-food/which-chains-have-most-loyal-customers/.

Lippman SA, McCardle KF (1997) The competitive newsboy. *Operations research* 45(1):54–65.

Liu Y, Yang L (2020) Order ahead for pickup: Promise or peril? *Available at SSRN* .

Maglaras C, Yao J, Zeevi A (2013) Optimal price and delay differentiation in queueing systems. *Available at SSRN 2297042* .

Maglaras C, Zeevi A (2003) Pricing and capacity sizing for systems with shared resources: Approximate solutions and scaling relations. *Management Science* 49(8):1018–1038.

Mahajan S, Van Ryzin G (2001) Inventory competition under dynamic consumer choice. *Operations research* 49(5):646–657.

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

Nageswaran L, Cho SH, Scheller-Wolf A (2020) Consumer return policies in omnichannel operations. *Management Science* .

Naor P (1969) The regulation of queue size by levying tolls. *Econometrica: journal of the Econometric Society* 15–24.

Nazerzadeh H, Randhawa RS (2018) Near-optimality of coarse service grades for customer differentiation in queueing systems. *Production and Operations Management* 27(3):578–595.

Netessine S, Shumsky RA (2005) Revenue management games: Horizontal and vertical competition. *Management Science* 51(5):813–831.

Plambeck EL (2004) Optimal leadtime differentiation via diffusion approximations. *Operations Research* 52(2):213–228.

Roet-Green R, Yuan Y (2019) Information visibility in omnichannel queues. *Available at SSRN* .

Van Mieghem JA (2000) Price and service discrimination in queuing systems: Incentive compatibility of gc μ scheduling. *Management Science* 46(9):1249–1267.

Veeraraghavan S, Debo L (2009) Joining longer queues: Information externalities in queue choice. *Manufacturing & Service Operations Management* 11(4):543–562.

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

Yahalom T, Harrison J, Kumar S (2006) Designing and pricing incentive compatible grades of service in queueing systems, technical report .

APPENDIX A

Proofs of Technical Results in Chapter 1

A.1. Proofs of Theorems

Proof of Theorem 1

First we solve problem (1.8) and characterize R_N^1. Applying Lemma 5 we have, $\overline{\lambda}_1 v_1 = \overline{\lambda}_2 v_2 = \cdots = \overline{\lambda}_N v_N$. Thus, we can write $\lambda_k v_k$ as $\lambda_k v_k = \overline{\lambda}_{k-1} v_{k-1} - \overline{\lambda}_{k-1} v_k$, which can be further written as $(\lambda_1 + \lambda_2 + \cdots + \lambda_{k-1})(v_{k-1} - v_k) = \frac{\lambda_1 v_1}{v_{k-1}}(v_{k-1} - v_k) = \lambda_1 v_1(1 - \frac{v_k}{v_{k-1}})$. Using this we can express the WCRR as $R_N^1 = \inf_{\lambda>0, v>0} \left\{ \frac{\lambda_1 v_1}{\lambda_1 v_1 + \lambda_2 v_2 + \cdots + \lambda_N v_N} \right\} = \inf_{\lambda>0, v>0} \left\{ \frac{\lambda_1 v_1}{\lambda_1 v_1 + \lambda_1 v_1 (1 - \frac{v_2}{v_1}) + \cdots + \lambda_1 v_1 (1 - \frac{v_N}{v_{N-1}})} \right\}$, which is equivalent to,

$$\inf_{v>0} \left\{ \frac{1}{N - \frac{v_2}{v_1} - \frac{v_3}{v_2} - \cdots - \frac{v_N}{v_{N-1}}} \right\} = \frac{1}{N}.$$

Thus, under the optimal solution to (1.8), the valuations are such that $\frac{v_k}{v_{k+1}} \to \infty$ for all $k = 1, 2, \ldots, N-1$. We also note that, the revenue for service grade k in the worst case optimal menu, $\lambda_k\big(v_k - h_k\big(\frac{v_{k-1} - v_k}{h_{k-1} - h_k}\big)\big)$, under the optimal solution becomes $\lambda_k v_k$. This implies that as $\frac{v_k}{v_{k+1}} \to \infty$ and $\frac{h_k}{h_{k+1}} \to \infty$, we have $\frac{v_k / v_{k+1}}{h_k / h_{k+1}} \to 0$ for all $k = 1, 2, \ldots, N-1$. Now, if the firm offers $K > 1$ service grades, the WCRR cannot be lower than it would be for $K = 1$ since the firm can always offer one service grade if it were optimal to do so. Therefore, $\frac{1}{N} = R_N^1 \leq R_N^K$. Moreover, using the upper bound $\overline{R}_N^K$, as established in (1.13), we have $R_N^K \leq \overline{R}_N^K = \frac{K}{N}$. Thus, we have $\frac{1}{N} \leq R_N^K \leq \frac{K}{N}$. ■

Proof of Theorem 2

The solution to problem (1.21) can be found equivalently by solving the following problem:

$$\min_{1\le i<j} \inf_{\delta_k} \left\{ \sum_{k=i}^{j-1} \left(\frac{1}{c_k(c_k+1)(\delta_k-1)} - \frac{1}{1+c_k} \right) \right\}$$
$$\text{s.t} \prod_{k=i}^{j-1} \delta_k \le \Delta_h,\ \delta_k > 1 + \frac{1}{c_k} \text{ for all } k = i, i+1, \ldots, j-1.$$

For fixed i and j, this is a convex constrained optimization problem over δ_k. The Lagrangian is given by, $\mathcal{L}(\delta_k, \mu) = \sum_{k=i}^{j-1} \left(\frac{1}{c_k(c_k+1)(\delta_k-1)} - \frac{1}{1+c_k} + \mu \log \delta_k \right) - \mu \log \Delta_h$. We do not include the price condition in the Lagrangian since it is a strict inequality. We ensure that the optimal solution satisfies price condition. Using the first order optimality (KKT) conditions, we get, $\nabla_{\delta_k} \mathcal{L} = -\frac{1}{c_k(c_k+1)(\delta_k-1)^2} + \frac{\mu}{\delta_k} = 0$, which is equivalent to

$$\frac{1}{c_k(c_k+1)(\delta_k-1)^2} = \frac{\mu}{\delta_k}. \tag{A.1}$$

Equation (A.1) is quadratic in δ_k. Solve for δ_k, and ensuring that the price condition holds, i.e. $\delta_k > 1 + \frac{1}{c_k}$ for all $k = i, i+1, \ldots, j-1$, we get,

$$\delta_k = \frac{2c_k\mu + (4c_k^2\mu + 4c_k\mu + 1)^{\frac{1}{2}} + 2c_k^2\mu + 1}{2c_k(c_k+1)\mu} \tag{A.2}$$

$$= 1 + \frac{1}{(1+c_k)\sqrt{\mu}} \left(1 + \frac{1}{c_k} + \frac{1}{4c_k^2\mu} \right) + \frac{1}{2c_k(c_k+1)\mu}. \tag{A.3}$$

The price condition implies that $\delta_k > 1$ for all $k = i, i+1, \ldots, j-1$. Thus, from equation (A.1) we note that the Lagrange multiplier, μ, is a strictly positive constant. This implies that the constraint, $\prod_{k=i}^{j-1} \delta_k \le \Delta_h$ is binding at optimality. Hence, the optimal values of

δ_k are such that

$$\prod_{k=i}^{j-1} \delta_k = \Delta_h. \tag{A.4}$$

Since $\{\lambda_k\}_{k=1}^{\infty}$ is a fixed sequence, $\{c_k\}_{k=1}^{\infty}$ is also fixed and is exogenously given in problem (1.21). Furthermore, our assumption of bounded arrival rates implies that $\{c_k\}_{k=1}^{\infty}$ is a diverging sequence. To see this, we note that c_k is defined as $c_k = \frac{\bar{\lambda}_k}{\lambda_{k+1}} = \frac{\sum_{n=1}^{k} \lambda_n/\lambda_1}{\lambda_{k+1}/\lambda_1}$. Assumption of bounded arrival rates implies that for some $m > 0$ and $M < \infty$, $m \leq \frac{\lambda_k}{\lambda_1} \leq M$ for all $k = 1, 2, \ldots, \infty$. Thus, using the bounds on the arrival rates, we have $\frac{m}{M}k \leq \frac{\sum_{n=1}^{k} \lambda_n/\lambda_1}{\lambda_{k+1}/\lambda_1} \leq \frac{M}{m}k$. Hence, $\frac{m}{M}k \leq c_k \leq \frac{M}{m}k$ for all $k = 1, 2, \ldots, \infty$. Before we show that the optimal solution to (1.21) is realized as $i \to \infty$, we continue with this assumption for now and solve the optimization problem. Since we have shown that $\{c_k\}_{k=1}^{\infty}$ is a diverging sequence, this implies that $c_k \to \infty$ as $i \to \infty$ for all $k = i, i+1, \ldots, j-1$. The right hand side in equation (A.1) is positive and finite, which further implies that $\delta_k \to 1$ for all $k = i, i+1, \ldots, j-1$. Consequently, from equation (A.4) this further implies that $(j-i) \to \infty$, i.e. the left hand side is an infinite product. Since there are infinitely many customer classes, it makes sense that the worst case scenario is realized as $(j-i) \to \infty$, i.e. infinitely many customer classes join service. Now, using (A.3) we can write δ_k as,

$$\delta_k = 1 + \epsilon_k = 1 + \frac{1}{(c_k+1)\sqrt{\mu}} + \mathcal{O}\left(\frac{1}{c_k^2}\right), \tag{A.5}$$

where $\mathcal{O}\left(\frac{1}{c_k^2}\right)$ represents the higher order terms. As $c_k \to \infty$, ignoring $\mathcal{O}\left(\frac{1}{c_k^2}\right)$, we get,

$$\frac{1}{(c_k+1)(\delta_k - 1)} = \sqrt{\mu}. \tag{A.6}$$

We can take logarithmic transform on both sides of the equations (A.4) to obtain

$$\sum_{k=i}^{j-1} \log(\delta_k) = \log(\Delta_h)$$

. Using the Taylor's approximation $\log(1+\epsilon_k) = \epsilon_k$ on $\log(\delta_k)$, using (A.5) we get,

$$\sum_{k=i}^{j-1} \log(1+\epsilon_k) = \sum_{k=i}^{j-1} \frac{1}{(c_k+1)\sqrt{\mu}} = \log(\Delta_h). \tag{A.7}$$

Now, using (A.6) we can rewrite the objective function as

$$\sum_{k=i}^{j-1} \left(\frac{1}{c_k(c_k+1)(\delta_k - 1)} - \frac{1}{1+c_k} \right) = \sum_{k=i}^{j-1} \left(\frac{\sqrt{\mu}}{c_k} - \frac{1}{1+c_k} \right). \tag{A.8}$$

Since $c_k \to \infty$ for all $k = i, i+1, \ldots, j-1$, we note that $\sum_{k=i}^{j-1} \frac{1}{1+c_k} = \sum_{k=i}^{j-1} \frac{1}{c_k\left(1+\frac{1}{c_k}\right)} = \sum_{k=i}^{j-1} \frac{1}{c_k}$. Denoting $\sum_{k=i}^{j-1} \frac{1}{1+c_k} = \sum_{k=i}^{j-1} \frac{1}{c_k}$ as S_{ij} and using (A.7) we have $\sqrt{\mu} = \dfrac{S_{ij}}{\log(\Delta_h)}$. Thus, we can rewrite (A.8) as

$$\sum_{k=i}^{j-1} \left(\frac{\sqrt{\mu}}{c_k} - \frac{1}{1+c_k} \right) = \frac{S_{ij}^2}{\log(\Delta_h)} - S_{ij}. \tag{A.9}$$

Now, we have already shown that $\frac{m}{M}k \leq c_k \leq \frac{M}{m}k$ for all $k = 1, 2, \ldots, \infty$. This implies $\frac{m}{M}\sum_{k=i}^{j-1} \frac{1}{k} \leq \sum_{k=i}^{j-1} \frac{1}{c_k} = S_{ij} \leq \frac{M}{m}\sum_{k=i}^{j-1} \frac{1}{k}$. Thus, S_{ij} is a diverging series. The value of S_{ij} that minimizes (A.9) is $S_{ij} = \frac{1}{2}\log(\Delta_h)$. Thus $\sqrt{\mu} = \frac{1}{2}$. The relationship between optimal values of the indices i^* and j^* is determined by the fact $S_{ij} = \frac{1}{2}\log(\Delta_h)$. Plugging S_{ij} in (A.9), we have $\sum_{k=i}^{j-1} \left(\dfrac{1}{1+c_k} - \dfrac{\sqrt{\mu}}{c_k} \right) = \frac{1}{4}\log(\Delta_h)$. Thus $R_\infty^1(\Delta_h)$ in (1.21) is given as, $R_\infty^1(\Delta_h) = \dfrac{1}{1+\frac{1}{4}\log(\Delta_h)}$. Now for the specific case of equal arrival rates, i.e., $\lambda_k = 1$ for all $k = 1, 2, \ldots, \infty$, we have $S_{i^*j^*} = \sum_{k=i^*}^{j^*-1} \frac{1}{k}$. Using logarithmic approximation for the

harmonic series, we get,

$$S_{i^*j^*} = (\log(j^*) + \gamma + e_{j^*}) - (\log(i^*) + \gamma + e_{i^*}) = \log\left(\frac{j^*}{i^*}\right) + (e_{j^*} - e_{i^*}),$$

where $e_{i^*} \to 0$ and $e_{j^*} \to 0$ as $i^* \to \infty$ and $j^* \to \infty$ respectively, and γ is the Euler-Mascheroni constant. Thus $S_{i^*j^*} = \log\left(\frac{j^*}{i^*}\right)$. We also note from (A.7) that, $S_{i^*j^*} = \sqrt{\mu}\log(\Delta_h) = \frac{1}{2}\log(\Delta_h)$. Equating the two expressions for $S_{i^*j^*}$ we have $\log\left(\frac{j^*}{i^*}\right) = \frac{1}{2}\log(\Delta_h)$ which implies $i^* = \frac{j^*}{\sqrt{\Delta_h}}$. In addition, using (A.5), we have $\delta_k^* = 1 + \frac{2}{k+1}$ for all k such that $i^* \leq k \leq j^* - 1$. Finally applying lemma 7, we show that indeed $i^* \to \infty$ under the optimal solution to problem (1.21). Consequently $(j^* - i^*) \to \infty$. This also proves that $R_N^1(\Delta_h) - R_\infty^1(\Delta_h) > 0$ for any finite N, and $R_N^1(\Delta_h) \to R_\infty^1(\Delta_h)$ as $N \to \infty$.

■

Proof of Theorem 3

The solution to problem (1.25) can be found by equivalently solving the following problem:

$$\inf_{\delta_k, c_k} \sum_{k=1}^{\infty} \left(\frac{1}{c_k(c_k+1)(\delta_k - 1)} - \frac{1}{1+c_k}\right) \tag{A.10}$$

$$\text{s.t. } \prod_{k=1}^{\infty} \delta_k \leq \Delta_h, \tag{A.11}$$

$$\prod_{k=1}^{\infty} \left(1 + \frac{1}{c_k}\right) \leq \Delta_v, \tag{A.12}$$

$$\delta_k > 1 + \frac{1}{c_k}, c_k > 0 \text{ for all } k = 1, 2, \ldots, \infty. \tag{A.13}$$

We begin by writing the Lagrangian for this problem without constraints (A.13) since these are strict inequalitites. We will verify that (A.13) satisfies under the optimal solution. Thus, we have, $\mathcal{L} = \sum_{k=1}^{\infty} \Big(\frac{1}{c_k(c_k+1)(\delta_k-1)} - \frac{1}{1+c_k} + \mu_h \log(\delta_k) + \mu_v \log\big(1 + \frac{1}{c_k}\big)\Big) - \mu_h \log(\Delta_h) - \mu_v \log(\Delta_v)$. Taking the first order optimality (KKT) conditions, we have $\nabla_{\delta_k}\mathcal{L} = 0$ and $\nabla_{c_k}\mathcal{L} = 0$, which imply

$$\mu_h = \frac{\delta_k}{c_k(c_k+1)(\delta_k-1)^2}, \text{ and } \mu_v = \frac{c_k}{1+c_k} - \frac{1+2c_k}{c_k(1+c_k)(\delta_k-1)} \tag{A.14}$$

respectively. Solving the above set of equations, (A.14), with constraints (A.13), i.e. for $\delta_k > 1 + \frac{1}{c_k}$ and $c_k > 0$, we get,

$$\delta_k = \frac{2c_k\mu_h + (4\mu_h c_k^2 + 4\mu_h c_k + 1)^{1/2} + 2c_k^2\mu_h + 1}{2\mu_h c_k(c_k+1)},$$

$$c_k = \frac{\delta_k\mu_v - \mu_v + (\delta_k^2\mu_v^2 - 2\delta_k\mu_v^2 + 4\delta_k + \mu_v^2)^{1/2} + 2}{2(\delta_k + \mu_v - \delta_k\mu_v - 1)}.$$

From (A.14), we note that $\mu_h > 0$ since the RHS is always positive. However, $\mu_v \geq 0$, i.e., it can be either zero or strictly positive. Applying Lemma 6, the optimal value of (A.10) is achieved as $c_k \to \infty$ for all $k = 1, 2, \ldots, \infty$. Using (A.14), this implies that, under the optimal solution, $\delta_k \to 1$ for all $k = 1, 2, \ldots, \infty$ since Lagrange multiplier μ_h is a finite positive quantity. Thus, ignoring higher order terms, we get

$$\delta_k = 1 + \frac{1}{(c_k+1)\sqrt{\mu_h}} \tag{A.15}$$

and

$$1 + \frac{1}{c_k} = 1 + \frac{1}{2}(1-\mu_v)(\delta_k - 1). \tag{A.16}$$

Combining (A.15) and (A.16), and the fact that $c_k \to \infty$ for all k, we get the following relation between the Lagrange multipliers μ_h and μ_v:

$$\sqrt{\mu_h} = \frac{1-\mu_v}{2}. \tag{A.17}$$

We can rearrange the expression for δ_k in (A.15) to obtain $\frac{\sqrt{\mu_h}}{c_k} = \frac{1}{c_k(1+c_k)(\delta_k-1)}$. Using the above relation and the fact that $c_k \to \infty$ for all k, we can rewrite the objective in (A.10) as follows:

$$\sum_{k=1}^{\infty}\Big(\frac{1}{c_k(c_k+1)(\delta_k-1)} - \frac{1}{1+c_k}\Big) = \sum_{k=1}^{\infty}\Big(\frac{\sqrt{\mu_h}}{c_k} - \frac{1}{1+c_k}\Big). \tag{A.18}$$

We note that $\sum_{k=1}^{\infty}\frac{1}{1+c_k} = \sum_{k=1}^{\infty}\frac{1}{c_k(1+\frac{1}{c_k})} = \sum_{k=1}^{\infty}\frac{1}{c_k}$ as $c_k \to \infty$ for all k. Thus, defining $S = \sum_{k=1}^{\infty}\frac{1}{1+c_k} = \sum_{k=1}^{\infty}\frac{1}{c_k}$, we can express (A.18) as

$$S\big(\sqrt{\mu_h} - 1\big). \tag{A.19}$$

Since $\mu_h > 0$, equation (A.11) binds at optimality. Replacing the value of δ_k in (A.11) using (A.15) we get, $\sum_{k=1}^{\infty}\log(1+\frac{1}{(1+c_k)\sqrt{\mu_h}}) = \log(\Delta_h)$. Since, $c_k \to \infty$, we can use the Taylor's approximation of $\log(1+\frac{1}{1+c_k}) = \frac{1}{1+c_k}$ to obtain the following expression:

$$\sum_{k=1}^{\infty}\frac{1}{1+c_k} = S = \sqrt{\mu_h}\log(\Delta_h). \tag{A.20}$$

If $\mu_v > 0$, then equation (A.12) binds at optimality and hence using (A.12) and (A.16) we get, $\sum_{k=1}^{\infty}\log(1+\frac{1}{2}(1-\mu_v)(\delta_k-1)) = \log(\Delta_v)$. Since $\delta_k \to 1$, we can use the Taylor's approximation of $\log(1+\frac{1}{2}(1-\mu_v)(\delta_k-1)) = \frac{1}{2}(1-\mu_v)(\delta_k-1)$ and obtain the following

expression: $\sum_{k=1}^{\infty}(\delta_k - 1) = \frac{2\log(\Delta_v)}{1-\mu_v}$. We can use (A.15) in this equation to obtain:

$$\sum_{k=1}^{\infty} \frac{1}{1+c_k} = S = \frac{2\log(\Delta_v)}{1-\mu_v}\sqrt{\mu_h}. \tag{A.21}$$

Combining (A.20) and (A.21) we get, $\log(\Delta_h) = \frac{2\log(\Delta_v)}{1-\mu_v}$. From this equation, we also note that $\mu_v > 0$ is equivalent to $\frac{2\log(\Delta_v)}{\log(\Delta_h)} < 1$ or $\Delta_v < \sqrt{\Delta_h}$. Combining this with (A.17), we get,

$$\sqrt{\mu_h} = \begin{cases} \dfrac{\log(\Delta_v)}{\log(\Delta_h)} & \mu_v > 0 \quad (\Delta_v < \sqrt{\Delta_h}), \\ \dfrac{1}{2} & \mu_v = 0 \quad (\Delta_v \geq \sqrt{\Delta_h}). \end{cases} \tag{A.22}$$

Consequently, combining (A.22) with (A.20) we get,

$$S = \begin{cases} \log(\Delta_v) & \Delta_v < \sqrt{\Delta_h}, \\ \frac{1}{2}\log(\Delta_h) & \Delta_v \geq \sqrt{\Delta_h}. \end{cases}$$

Finally, replacing S using the above expression in (A.19), the optimal solution to (1.25) is given by,

$$R^1_{\infty}(\Delta_h, \Delta_v) = \begin{cases} \dfrac{1}{1 + \dfrac{\log(\Delta_v)}{\log(\Delta_h)} \log\left(\dfrac{\Delta_h}{\Delta_v}\right)} & \Delta_v < \sqrt{\Delta_h}, \\ \dfrac{1}{1+\frac{1}{4}\log(\Delta_h)} & \Delta_v \geq \sqrt{\Delta_h}. \end{cases}$$

■

A.2. Proofs of Propositions

Proof of Proposition 1

A menu in set $\mathbb{M}$ induces some customer segmentation σ. This implies that customer class i_{k-1} joins service grade (p_{k-1}, d_{k-1}) for $k > 1$. Incentive compatibility ensures that customer class i_{k-1} should have non-positive utility by joining service grade k, which implies,

$$v_{i_{k-1}} - p_k - h_{i_{k-1}} d_k \leq 0. \tag{A.23}$$

Using Definition 2, since the menu belongs to set $\mathbb{M}$, we have, $v_{i_{k-1}} - p_{k-1} - h_{i_{k-1}} d_{k-1} = 0$. Subtracting this equation from (A.23), we have $(p_{k-1} - p_k) + h_{i_{k-1}}(d_{k-1} - d_k) \leq 0$. Since we index service grades in decreasing order of their prices, $p_{k-1} > p_k$, which implies $d_k > d_{k-1}$. Again, since the menu belongs to set $\mathbb{M}$, from Definition 2 we have, $v_{i_k} - p_k - h_{i_k} d_k = 0$. Subtracting this equation from (A.23), we have $(v_{i_{k-1}} - v_{i_k}) + (h_{i_k} - h_{i_{k-1}}) d_k \leq 0$. Since $v_{i_{k-1}} > v_{i_k}$, this implies $h_{i_{k-1}} > h_{i_k}$. This is intuitive, otherwise delay differentiation wouldn't have been possible. Using equation $v_{i_k} - p_k - h_{i_k} d_k = 0$ we can replace d_k in (A.23) to obtain $v_{i_{k-1}} - \frac{h_{i_{k-1}}}{h_{i_k}} v_{i_k} + p_k \left(\frac{h_{i_{k-1}}}{h_{i_k}} - 1 \right) \leq 0$. As $h_{i_{k-1}} > h_{i_k}$, the second term in this inequality is positive, and we note that a revenue-maximizing menu would increase p_k until the inequality binds. Therefore, the service grade (p_k, d_k) is characterized by the following two equations:

$$v_{i_{k-1}} - p_k - h_{i_{k-1}} d_k = 0 \text{ and } v_{i_k} - p_k - h_{i_k} d_k = 0. \tag{A.24}$$

Solving these two equations we get $p_k = v_{i_k} - h_{i_k}\left(\frac{v_{i_{k-1}} - v_{i_k}}{h_{i_{k-1}} - h_{i_k}}\right)$ and $d_k = \left(\frac{v_{i_{k-1}} - v_{i_k}}{h_{i_{k-1}} - h_{i_k}}\right)$. Now, for $k = 1$, all the customer classes $1, 2, \ldots, i_1$ join service grade (p_1, d_1). Also, from Definition 2 we have $v_{i_1} - p_1 - h_{i_1} d_1 = 0$. As a result, we can conclude that the price p_1 is maximized if the delay $d_1 = 0$ (with Assumption 1), i.e., $p_1 = v_{i_1}$. This proves the second part of the Proposition.

We will prove the first part of the Proposition by induction. First, let's suppose $k = 1$. We know that classes $1, 2, \ldots, i_1$ join grade $(p_1, d_1) = (v_{i_1}, 0)$ and classes $i_1+1, i_1+2, \ldots, i_2$ join grade (p_2, d_2). Let's consider an alternative menu which offers service grade $(\hat{p} = v_{i_2}, \hat{d} = 0)$ instead of grades (p_1, d_1) and (p_2, d_2) leaving all other service grades in the menu unchanged. This alternative menu results in a new customer segmentation where customer classes $1, 2, \ldots, i_2$ to join $(\hat{p}, \hat{d})$, and the choice of service grade for the other classes do not change. Since the revenue-maximizing menu offers differentiated service grades (p_1, d_1) and (p_2, d_2), offering $(\hat{p}, \hat{d})$ instead would lead to a lower revenue. This implies, $\overline{\lambda}_{i_2}\hat{p} < \overline{\lambda}_{i_1} p_1 + (\overline{\lambda}_{i_2} - \overline{\lambda}_{i_1}) p_2$ which is equivalent to

$$\overline{\lambda}_{i_2}(\hat{p} - p_2) < \overline{\lambda}_{i_1}(p_1 - p_2) \tag{A.25}$$

where $\overline{\lambda}_l = \sum_{m=1}^{l} \lambda_m$. Using equations (A.24) corresponding to $k = 1$ and $k = 2$, we can derive the following relation between p_1 and p_2: $p_1 = p_2 + h_{i_1}(d_2 - d_1)$. Using Definition 2, service grade $(\hat{p}, \hat{d})$ is chosen such that $v_{i_2} - \hat{p} - h_{i_2}\hat{d} = 0$. Additionally, from (A.24) we have $v_{i_2} - p_2 - h_{i_2} d_2 = 0$. Combining these two equations, we get $\hat{p} = p_2 + h_{i_2}(d_2 - \hat{d})$. Combined with $d_1 = 0$ and $\hat{d} = 0$ we have the following two relations: $p_1 - p_2 = h_{i_1} d_2$ and $\hat{p} - p_2 = h_{i_2} d_2$. Replacing these expressions in (A.25) we get, $\frac{h_{i_1}}{h_{i_2}} > \frac{\overline{\lambda}_{i_2}}{\overline{\lambda}_{i_1}}$.

Now, let's suppose $k > 1$. We consider service grades (p_k, d_k) and (p_{k+1}, d_{k+1}) on the revenue-maximizing menu. Similar to the argument for $k = 1$, let's consider an alternative menu which offers service grade $(\hat{p}, \hat{d})$ instead of grades (p_k, d_k) and (p_{k+1}, d_{k+1}) leaving all other grades in the menu unchanged such that customer classes $i_{k-1}+1, i_{k-1}+2, \ldots, i_{k+1}$ join $(\hat{p}, \hat{d})$ and the choice of service grade for the other classes do not change. Service grade $(\hat{p}, \hat{d})$ is chosen in such a way that it satisfies incentive compatibility constraints and results in the highest possible revenue corresponding to the new customer segmentation induced by the alternative menu. Thus, ensuring appropriate incentive compatibility and individual rationality constraints hold, service grade $(\hat{p}, \hat{d})$ is chosen such that it is related to the service grades (p_k, d_k) and (p_{k+1}, d_{k+1}) in the original menu, through the following equations:

$$p_k = \hat{p} + h_{i_{k-1}}(\hat{d} - d_k), \tag{A.26}$$

$$\hat{p} = p_{k+1} + h_{i_{k+1}}(d_{k+1} - \hat{d}). \tag{A.27}$$

Moreover, service grades (p_k, d_k) and (p_{k+1}, d_{k+1}) are related through the following equation which is derived using the equations (A.24) for indices k and $k+1$:

$$p_k = p_{k+1} + h_{i_k}(d_{k+1} - d_k). \tag{A.28}$$

Subtracting (A.26) from (A.28), we get $\hat{p} - p_{k+1} = h_{i_k}(d_{k+1} - d_k) - h_{i_{k-1}}(\hat{d} - d_k)$. Replacing d_{k+1} in this equation using (A.27) we get the following expression:

$$\frac{\hat{p} - p_{k+1}}{\hat{d} - d_k} = \frac{h_{i_k} - h_{i_{k-1}}}{1 - \frac{h_{i_k}}{h_{i_{k+1}}}}. \tag{A.29}$$

Since the revenue-maximizing menu offers delay differentiated grades (p_k, d_k) and (p_{k+1}, d_{k+1}), offering $(\hat{p}, \hat{d})$ instead would lead to a lower revenue. This implies, $(\overline{\lambda}_{i_k} - \overline{\lambda}_{i_{k-1}})p_k + (\overline{\lambda}_{i_{k+1}} - \overline{\lambda}_{i_k})p_{k+1} > (\overline{\lambda}_{i_{k+1}} - \overline{\lambda}_{i_{k-1}})\hat{p}$. In this inequality, we can replace p_k using (A.26) to get $(\overline{\lambda}_{i_k} - \overline{\lambda}_{i_{k-1}})(\hat{p} + h_{i_{k-1}}(\hat{d} - d_k)) + (\overline{\lambda}_{i_{k+1}} - \overline{\lambda}_{i_k})p_{k+1} > (\overline{\lambda}_{i_{k+1}} - \overline{\lambda}_{i_{k-1}})\hat{p}$. We use (A.29) to replace $(\hat{d} - d_k)$ in this inequality to obtain

$$(\overline{\lambda}_{i_k} - \overline{\lambda}_{i_{k-1}})\left(\hat{p} + h_{i_{k-1}}\left(\frac{\hat{p} - p_{k+1}}{h_{i_k} - h_{i_{k-1}}}\right)\left(1 - \frac{h_{i_k}}{h_{i_{k+1}}}\right)\right) + (\overline{\lambda}_{i_{k+1}} - \overline{\lambda}_{i_k})p_{k+1} > (\overline{\lambda}_{i_{k+1}} - \overline{\lambda}_{i_{k-1}})\hat{p}.$$

Dividing both sides by $(\overline{\lambda}_{i_k} - \overline{\lambda}_{i_{k-1}})$ and rearranging the inequality we can cancel $\hat{p} - p_{k+1}$ from both sides to obtain, $\left(\frac{h_{i_{k-1}}}{h_{i_k} - h_{i_{k-1}}}\right)\left(1 - \frac{h_{i_k}}{h_{i_{k+1}}}\right) > \left(\frac{\overline{\lambda}_{i_{k+1}} - \overline{\lambda}_{i_k}}{\overline{\lambda}_{i_k} - \overline{\lambda}_{i_{k-1}}}\right)$ which is equivalent to

$$\left(\frac{h_{i_k}}{h_{i_{k+1}}} - 1\right) > \left(\frac{\overline{\lambda}_{i_{k+1}} - \overline{\lambda}_{i_k}}{\overline{\lambda}_{i_k} - \overline{\lambda}_{i_{k-1}}}\right)\left(1 - \frac{h_{i_k}}{h_{i_{k-1}}}\right). \tag{A.30}$$

By induction argument, first we assume that the result holds for $k-1$, i.e. $\frac{h_{i_{k-1}}}{h_{i_k}} > \frac{\overline{\lambda}_{i_k}}{\overline{\lambda}_{i_{k-1}}}$. Now, combining this with inequality (A.30) we get, $\frac{h_{i_k}}{h_{i_{k+1}}} > \frac{\overline{\lambda}_{i_{k+1}}}{\overline{\lambda}_{i_k}}$, which shows that the result holds for k. This concludes the induction proof. Finally, for any two service grades (p_j, d_j) and (p_k, d_k) in the menu, we can write $\frac{h_{i_j}}{h_{i_k}} = \frac{h_{i_j}}{h_{i_j+1}}\frac{h_{i_j+1}}{h_{i_j+2}} \cdots \frac{h_{i_k-1}}{h_{i_k}} > \frac{\overline{\lambda}_{i_j+1}}{\overline{\lambda}_{i_j}}\frac{\overline{\lambda}_{i_j+2}}{\overline{\lambda}_{i_j+1}} \cdots \frac{\overline{\lambda}_{i_k}}{\overline{\lambda}_{i_k-1}} = \frac{\overline{\lambda}_{i_k}}{\overline{\lambda}_{i_j}}$. ■

Proof of Proposition 2

From (1.15) and (1.16) we have $\frac{1}{1+\sum_{k=2}^{N}\frac{M}{1+(k-1)m}} = R_N^1 \leq R_N^K \leq \frac{K}{1+\sum_{k=2}^{N}\frac{M}{1+(k-1)m}}$. The arrival rates are such that $0 < m \leq \frac{\lambda_k}{\lambda_1} \leq M < \infty$ for all $k = 1, 2, \ldots, N$. This implies

$0 < m \leq 1 \leq M < \infty$. Using m and M, we can come up with the following lower bound:

$$\frac{1}{1+\sum_{k=2}^{N}\frac{M}{1+(k-1)m}} \geq \frac{1}{\frac{M}{m}+\sum_{k=2}^{N}\frac{M}{m+(k-1)m}} = \frac{1}{\frac{M}{m}\sum_{k=1}^{N}\frac{1}{k}}.$$

Similarly, using m and M, we can come up with the following upper bound:

$$\frac{K}{1+\sum_{k=2}^{N}\frac{M}{1+(k-1)m}} \leq \frac{K}{1+\sum_{k=2}^{N}\frac{M}{M+(k-1)M}} = \frac{1}{\sum_{k=1}^{N}\frac{1}{k}}.$$

We can use the logarithmic approximation for the harmonic series to write $\sum_{k=1}^{N}\frac{1}{k} = \log(N)+\gamma_N$ where $\gamma < \gamma_N < \gamma+\frac{1}{2}, \lim_{N\to\infty}\gamma_N \to \gamma$ and γ is the Euler-Mascheroni constant. This proves the result. ■

Proof of Proposition 3

Optimization problem (1.19) can be rewritten as,

$$\inf_{v_k>0,\delta_k,z} z$$

$$\text{s.t. } z \geq \frac{\overline{\lambda}_k v_k}{\overline{\lambda}_i v_i + \lambda_{i+1}v_{i+1}\left(\frac{\delta_i - \frac{v_i}{v_{i+1}}}{\delta_i - 1}\right) + \cdots + \lambda_j v_j \left(\frac{\delta_{j-1} - \frac{v_{j-1}}{v_j}}{\delta_{j-1}-1}\right)}, \text{ for all } k = 1, 2, \ldots, N,$$

$$\log(\Delta_h) \geq \sum_{k=1}^{N-1}\log(\delta_k),$$

$$\delta_k \geq 1,\ v_k > v_{k+1} \text{ for all } k = 1, 2, \ldots, N-1,$$

$$\delta_k > \frac{v_k}{v_{k+1}} \text{ for all } k = i, i+1, \ldots, j-1.$$

We note that constraints $\delta_k \geq 1$ would not be binding for all k where $i \leq k \leq j-1$. Hence considering all the weak inequalities, the Lagrangian for this problem can be written

as,

$$\mathcal{L} = z(1-\alpha_1-\alpha_2-\cdots-\alpha_N) + \frac{\alpha_1\overline{\lambda}_1 v_1 + \alpha_2\overline{\lambda}_2 v_2 + \cdots + \alpha_N\overline{\lambda}_N v_N}{\overline{\lambda}_i v_i + \lambda_{i+1}v_{i+1}\left(\frac{\delta_i - \frac{v_i}{v_{i+1}}}{\delta_i - 1}\right) + \cdots + \lambda_j v_j\left(\frac{\delta_{j-1} - \frac{v_{j-1}}{v_j}}{\delta_{j-1}-1}\right)}$$
$$+ \mu(\sum_{k=1}^{N-1}\log(\delta_k) - \log(\Delta_h)) + \sum_{k\notin\{i,i+1,\ldots,j-1\}}\beta_k(1-\delta_k)$$

where α_k, β_k and μ are the Lagrange multipliers. Taking the first order KKT conditions, we get,

$$\nabla_z\mathcal{L} = 1-\alpha_1-\alpha_2-\cdots-\alpha_N = 0 \tag{A.31}$$

which implies that not all α_k are zero. Now, for all k such that $1 \le k \le i-1$ and $j+1 \le k \le N$, we have, $\nabla_{v_k}\mathcal{L} = \dfrac{\alpha_k\overline{\lambda}_k}{\overline{\lambda}_i v_i + \lambda_{i+1}v_{i+1}\left(\frac{\delta_i - \frac{v_i}{v_{i+1}}}{\delta_i - 1}\right) + \cdots + \lambda_j v_j\left(\frac{\delta_{j-1} - \frac{v_{j-1}}{v_j}}{\delta_{j-1}-1}\right)} = 0$, which implies that $\alpha_k = 0$ for all $1 \le k \le i-1$ and $j+1 \le k \le N$. Moreover, for all $i \le k \le j$, $\nabla_{v_k}\mathcal{L} = 0$ implies

$$\alpha_k\overline{\lambda}_k = \frac{A}{B}\frac{\partial B}{\partial v_k} \text{ where} \tag{A.32}$$

$$A = \alpha_1\overline{\lambda}_1 v_1 + \alpha_2\overline{\lambda}_2 v_2 + \cdots + \alpha_N\overline{\lambda}_N v_N \text{ and} \tag{A.33}$$

$$B = \overline{\lambda}_i v_i + \lambda_{i+1}v_{i+1}\left(\frac{\delta_i - \frac{v_i}{v_{i+1}}}{\delta_i - 1}\right) + \cdots + \lambda_j v_j\left(\frac{\delta_{j-1} - \frac{v_{j-1}}{v_j}}{\delta_{j-1}-1}\right). \tag{A.34}$$

By dual feasibility conditions, $\alpha_k \ge 0$ for all $k = 1, 2, \ldots, N$. Furthermore, B is the same as $\pi(\mathcal{M}_{ij})$ which is the revenue under the optimal menu. Revenue $\pi(\mathcal{M}_{ij})$ is comprised of the prices of service grades $i, i+1, \ldots, j$, which are determined by $v_i, v_{i+1}, \ldots, v_j$.

Perturbing any of these valuations would affect the prices of the service grades and hence change the revenue $\pi(\mathcal{M}_{ij})$. Hence, $\frac{\partial B}{\partial v_k} = \frac{\partial \pi(\mathcal{M}_{ij})}{\partial v_k} \neq 0$ for all k where $i \leq k \leq j$. Combining this with the fact that not all α_k are zero (from (A.31)) we have $A \neq 0$ in (A.33). This implies $\alpha_k > 0$ in (A.32). Thus, we have $\alpha_k = 0$ for all k where $1 \leq k \leq i-1$ and $j+1 \leq k \leq N$ and $\alpha_k > 0$ for all $i \leq k \leq j$. By complementary slackness condition, all constraints corresponding to positive Lagrange multipliers, i.e. $\alpha_k > 0$, would be strictly binding at optimality which implies that $\overline{\lambda}_i v_i = \overline{\lambda}_{i+1} v_{i+1} = \cdots = \overline{\lambda}_j v_j$. Hence we have $\frac{v_k}{v_{k+1}} = \frac{\overline{\lambda}_{k+1}}{\overline{\lambda}_k}$ for all k where $i \leq k \leq j-1$. Moreover, using $\alpha_1+\alpha_2+\cdots+\alpha_N = 1$ from (A.31), and the fact that $\overline{\lambda}_i v_i = \overline{\lambda}_{i+1} v_{i+1} = \cdots = \overline{\lambda}_j v_j$, we have from (A.33), $v_k = \frac{A}{\overline{\lambda}_k}$ for all $k = i, i+1, \ldots, j$. Constraints corresponding to $\alpha_k = 0$ are slack (or weakly binding). Hence we have, $1 \leq \frac{v_k}{v_{k+1}} \leq \frac{\overline{\lambda}_{k+1}}{\overline{\lambda}_k}$ for all k where $1 \leq k \leq i-1$ and $\frac{v_k}{v_{k+1}} \geq \frac{\overline{\lambda}_{k+1}}{\overline{\lambda}_k}$ for all $j \leq k \leq N-1$.

Finally, taking derivatives of the Lagrangian with respect to δ_k yields

$$\nabla_{\delta_k}\mathcal{L} = \begin{cases} \dfrac{\lambda_{k+1}(v_{k+1}-v_k)}{(\delta_k-1)^2}\dfrac{A}{B^2} + \dfrac{\mu}{\delta_k} & i \leq k \leq j-1, \\ \dfrac{\mu}{\delta_k} - \beta_k & 1 \leq k \leq i-1 \text{ and } j \leq k \leq N-1. \end{cases}$$

where A and B are defined in (A.33) and (A.34). For the case $i \leq k \leq j-1$, using $v_k = \frac{A}{\overline{\lambda}_k}$, $\nabla_{\delta_k}\mathcal{L} = 0$ implies $\mu = \dfrac{\delta_k}{c_k(c_k+1)(\delta_k-1)^2}.\dfrac{A^2}{B^2}$ where $c_k = \dfrac{\overline{\lambda}_k}{\lambda_{k+1}}$. This implies $\mu > 0$. Hence, by complementary slackness condition, the constraint $\prod_{k=1}^{N-1}\delta_k \leq \Delta_h$ would be binding at optimality which implies $\prod_{k=1}^{N-1}\delta_k = \Delta_h$. For the case where $1 \leq k \leq i-1$ and $j \leq k \leq N-1$, $\nabla_{\delta_k}\mathcal{L} = 0$ implies $\beta_k = \dfrac{\mu}{\delta_k}$. This implies $\beta_k > 0$ since $\mu > 0$. Again, by complementary slackness, the constraints $\delta_k \geq 1$ will be binding at optimality, i.e. $\delta_k = 1$ for all $1 \leq k \leq i-1$ and $j \leq k \leq N-1$. Combining all of

these we have, $\prod_{k=i}^{j-1}\delta_k = \Delta_h$ and $\dfrac{\delta_i}{c_i(c_i+1)(\delta_i-1)^2} = \dfrac{\delta_{i+1}}{c_{i+1}(c_{i+1}+1)(\delta_{i+1}-1)^2} = \cdots = \dfrac{\delta_{j-1}}{c_{j-1}(c_{j-1}+1)(\delta_{j-1}-1)^2}$. ■

A.3. Proofs of Lemmas and Corollary

Proof of Lemma 1

The WCRR for a firm offering $K = 1$ service grade, when there are N customer classes, is given by,

$$R_N^1 = \inf_{v>0,h>0}\left\{\frac{\pi_N^1}{\pi_N^*}\right\}.$$

Let's suppose, that the worst case optimal menu, WCOM, offers L service grades. Since π_N^* denotes the revenue corresponding to a revenue-maximizing menu, for each service grade l in the menu, there is at least one customer class among all the customer classes who join service grade l whose utility is zero. Let us denote this customer class as (v_{i_l}, h_{i_l}), where $l = 1, 2, \ldots, L$. This implies, $u_l : v_{i_l} - p_l - h_{i_l}d_l = 0$. Thus, we can characterize the worst case optimal menu by calculating prices and delays, (p_l, d_l), by solving the equations $u_l = 0$, for $l = 1, 2, \ldots, L$. As a result, π_N^* would be a function of only (v_{i_l}, h_{i_l}) for $l = 1, 2, \ldots, L$. On the other hand, $\pi_N^1 = \max\{\bar{\lambda}_1 v_1, \bar{\lambda}_2 v_2, \ldots, \bar{\lambda}_N v_N\}$, is a non-decreasing function of $v_1, v_2, \ldots, v_N$, where $\bar{\lambda}_k = \sum_{i=1}^{k}\lambda_i$. Now, let us consider a customer class k with (v_k, h_k), which is different from (v_{i_l}, h_{i_l}), that joins service grade l. We argue that, under the solution of (1.3), customer class k also has a utility zero, i.e., $v_k - p_l - h_k d_l = 0$. To see this, we note that if this wasn't true then the value of v_k could be reduced by an infinitesimally small amount to reduce WCRR since we have established that v_k doesn't appear in the denominator, π_N^*, and the numerator, π_N^1, decreases as v_k

decreases. Therefore, the worst case analysis ensures that the utility for each customer class is zero under the worst case optimal menu. Therefore, without loss of generality, we refer to class i_l as the customer class joining service grade l, with zero utility and having the least valuation among all the customer classes joining service grade l.

Let us consider service grades l and $l-1$. Customer classes i_l and i_{l-1} would satisfy the following incentive-compatibility constraints:

$$IC1 : v_{i_{l-1}} - p_l - h_{i_{l-1}} d_l < v_{i_{l-1}} - p_{l-1} - h_{i_{l-1}} d_{l-1},$$

$$IC2 : v_{i_l} - p_{l-1} - h_{i_l} d_{l-1} < v_{i_l} - p_l - h_{i_l} d_l,$$

These two ICs combined give us the following inequalities:

$$h_{i_l}(d_l - d_{l-1}) < p_{l-1} - p_l < h_{i_{l-1}}(d_l - d_{l-1}) \tag{A.35}$$

Since service grades are indexed in decreasing order of their prices, we have $p_{l-1} > p_l$ for all $l = 2, 3, \ldots, L$. Then, equation (A.35) implies $d_l > d_{l-1}$ and $h_{i_l} < h_{i_{l-1}}$. We also know that, $v_{i_{l-1}} - p_{l-1} - h_{i_{l-1}} d_{l-1} = 0,$ and $v_{i_l} - p_l - h_{i_l} d_l = 0$. Combining these equations with the IC2, we get the following:

$$v_{i_{l-1}} - p_{l-1} - h_{i_{l-1}} d_{l-1} = 0, \tag{A.36}$$

$$v_{i_l} - p_{l-1} - h_{i_l} d_{l-1} < 0. \tag{A.37}$$

Subtracting (A.36) from (A.37), we get,

$$v_{i_l} - v_{i_{l-1}} < (h_{i_l} - h_{i_{l-1}}) d_{l-1}. \tag{A.38}$$

We have already established that $h_{i_l} < h_{i_{l-1}}$, which, combined with (A.38) implies $v_{i_{l-1}} > v_{i_l}$. This implies, $i_{l-1} < i_l$ for all $l = 2, 3, \ldots, L$, since customer classes are indexed in decreasing order of their valuations. Up to this point we have established the following: (i) $p_1 > p_2 > \cdots > p_L$, (ii) $d_1 < d_2 < \cdots < d_L$, (iii) $v_{i_1} > v_{i_2} > \cdots > v_{i_L}$ and (iv) $h_{i_1} > h_{i_2} > \cdots > h_{i_L}$. What remains to be shown is that, any customer class whose valuation lies between $v_{i_{l-1}}$ and v_{i_l}, joins service grade l. To this end, let us consider a customer class (v, h), such that $v_{i_{l-1}} > v > v_{i_l}$. By definition, class i_{l-1} is the class with least valuation to join service grade $l-1$. This implies, no customer class with a lower valuation than $v_{i_{l-1}}$ joins any service service grade k where $k \leq l-1$. Thus, in order to show that, the class with valuation v joins grade l we need to show that it does not join any grade $k > l$. We will prove this by contradiction. Let's suppose, service grade (p_k, d_k) is such that $k > l$, i.e., $p_k < p_l$ and $d_k > d_l$. Let us assume that the class with valuation v joins grade (p_k, d_k). Since we have established that all customer classes, irrespective of which service grades they join, have zero consumer surplus under the worst case optimal menu, we have,

$$v - p_k - hd_k = 0. \tag{A.39}$$

Moreover, incentive compatibility would require this class to incur negative utility from joining service grade l. This implies,

$$v - p_l - hd_l < 0. \tag{A.40}$$

Also, we use the fact that, customer class i_l joins service grade l, which implies,

$$v_{i_l} - p_l - h_{i_l} d_l = 0. \tag{A.41}$$

We can rewrite (A.41) as, $v_{i_l} - p_l - h_{i_l} d_l - p_k + p_k - h_{i_l} d_k + h_{i_l} d_k = 0$. Rearranging the terms, we get,

$$\underbrace{(v_{i_l} - p_k - h_{i_l} d_k)}_{Term1} + \underbrace{(p_k - p_l + h_{i_l} d_k - h_{i_l} d_l)}_{Term2} = 0. \tag{A.42}$$

First, subtracting (A.41) from (A.40), we get, $(v - v_{i_l}) + d_l(h_{i_l} - h) < 0$. Since, $v > v_{i_l}$, this implies

$$h_{i_l} < h. \tag{A.43}$$

Secondly, subtracting (A.39) from (A.40), we get,

$$p_k - p_l + h(d_k - d_l) < 0. \tag{A.44}$$

Since, $d_k > d_l$ and $h_{i_l} < h$ from (A.43), replacing h by h_{i_l} in (A.44) we get,

$$p_k - p_l + h_{i_l}(d_k - d_l) < 0. \tag{A.45}$$

The left hand side of (A.45) is the same as *Term2* in (A.42). This implies that *Term1* in (A.42) is positive, i.e., $(v_{i_l} - p_k - h_{i_l} d_k) > 0$, which is a contradiction because it violates incentive compatibility by allowing a positive utility for customer class i_l by joining service grade k. Hence, any customer class with valuation v such that $v_{i_{l-1}} > v > v_{i_l}$, joins service

grade l. We have, thus, shown that the worst case menu induces customer segmentation σ as per Definition 1 and therefore it belongs to set $\mathbb{M}$ of menus as per Definition 2. ■

Proof of Lemma 2

In the presence of $N = 2$ customer classes, a revenue-maximizing firm could offer a menu which would result in either: i) both customer classes joining the same service grade, or, ii) two customer classes joining separate service grades, or, iii) only the higher customer class joining service and the lower customer class not joining service. In the first case, we say that the customer classes are *pooled* in the same service grade and in the second case we say that the customer classes are *differentiated* into separate service grades. For given valuations, delay sensitivities and the arrival rates, either of these possibilities could be realized under the optimal (revenue maximizing) menu. Without loss of generality, we refer to the higher valuation customer class as class 1 and the lower valuation class as class 2.

First, lets assume that a *pooling* menu is optimal. So, the firm offers a single service grade, (p, d), which is accepted by both customer classes, and the resulting revenue $(\lambda_1 + \lambda_2)p$ is maximized. The *individual rationality* constraints corresponding to the two customer classes, i.e. $v_1 - p - h_1 d \geq 0$ and $v_2 - p - h_2 d \geq 0$ need to hold. By Assumption 1, the firm could offer a service grade with $d = 0$, which allows for maximum possible price, p. Since by our convention, $v_1 > v_2$, this implies $p = v_2$ is the maximum possible price that allows both customer classes to join service. So the *pooling* service grade is $(p = v_2, d = 0)$ and the associated maximum revenue is $(\lambda_1 + \lambda_2)v_2$.

Next, let's assume that a *differentiating* menu is optimal. This implies, that the firm offers two service grades and charges price p_1 for the higher valuation customer class 1 and charges a price p_2 for the lower valuation customer class 2. The following *individual rationality* (IR) constraints,

$$\text{IR1: } v_1 - p_1 - h_1 d_1 \geq 0, \tag{A.46}$$

$$\text{IR2: } v_2 - p_2 - h_2 d_2 \geq 0, \tag{A.47}$$

and the two *incentive compatibility* (IC) constraints,

$$\text{IC1: } v_1 - p_1 - h_1 d_1 \geq v_1 - p_2 - h_1 d_2, \tag{A.48}$$

$$\text{IC2: } v_2 - p_2 - h_2 d_2 \geq v_2 - p_1 - h_2 d_1 \tag{A.49}$$

need to hold true. The objective function for the firm is to maximize $\lambda_1 p_1 + \lambda_2 p_2$. We note that, starting from any feasible value of p_1 and p_2 that sastify IR and the IC constraints, we could continue to increase both p_1 and p_2 by the same amount, and thus increasing the objective value until one of the IR binds while ensuring the ICs continue to hold. Say, (A.47) binds, we could next continue to increase p_1 until either (A.46) or (A.48) binds. Let's say IC1 (A.48) and IR2 (A.47) bind. We will show that this results in IR1 (A.46) binding at optimality as well. First, IR2 binding implies,

$$p_2 = v_2 - h_2 d_2. \tag{A.50}$$

Next, IC1 binding implies,

$$p_1 = p_2 + h_1(d_2 - d_1). \tag{A.51}$$

Now IC2 (A.49) combined with (A.50) implies, $0 \geq v_2 - p_1 - h_2 d_1$. First, substituting p_2 from (A.50) in (A.51) and then using it to replace p_1 in (A.49), and using (A.50) to substitute p_2 in (A.49), we get, $v_2 - (v_2 - h_2 d_2 + h_1(d_2 - d_1)) - h_2 d_1 \leq 0$, which implies

$$(h_1 - h_2)(d_1 - d_2) \leq 0. \tag{A.52}$$

Finally, from IR1 (A.46) we have, $v_1 - p_1 - h_1 d_1 \geq 0$. Substituting p_1 and p_2 using (A.50) and (A.51) in IR1, we get, $v_1 - (v_2 - h_2 d_2 + h_1(d_2 - d_1)) - h_1 d_1 \geq 0$ which implies,

$$\frac{v_1 - v_2}{h_1 - h_2} \geq d_2. \tag{A.53}$$

Substituting p_1 and p_2 using (A.50) and (A.51) in the objective function $(\lambda_1 p_1 + \lambda_2 p_2)$, we get,

$$(\lambda_1 p_1 + \lambda_2 p_2) = v_2(\lambda_1 + \lambda_2) - \lambda_1 h_1 d_1 + d_2(\lambda_1 h_1 - \lambda_2 h_2 - \lambda_1 h_2). \tag{A.54}$$

We note that the optimal revenue depends on the expression $(\lambda_1 h_1 - \lambda_2 h_2 - \lambda_1 h_2)$ in (A.54). If $\lambda_1 h_1 - \lambda_2 h_2 - \lambda_1 h_2 > 0$, which implies, $\frac{h_1}{h_2} > 1 + \frac{\lambda_2}{\lambda_1}$, then maximum revenue for the *differentiated* menu corresponds to $d_1 = 0$ and maximum feasible value of d_2 which is $d_2 = \frac{v_1 - v_2}{h_1 - h_2}$ from (A.53) which implies that IR1 binds. Also (A.52) holds which implies that these values of d_1 and d_2 are consistent with IC2. Since $p_2 = v_2 - h_2 \dfrac{v_1 - v_2}{h_1 - h_2}$ and $p_1 = p_2 + h_1(d_2 - d_1) = v_2 - (h_2 - h_1)\dfrac{v_1 - v_2}{h_1 - h_2} = v_1$, the optimal revenue for this case is

given by, $\lambda_1 v_1 + \lambda_2\Big(v_2 - h_2\frac{v_1-v_2}{h_1-h_2}\Big)$. Hence the *differentiating* service grades are given by,

$$(p_1, d_1) : (v_1, 0) \text{ and } (p_2, d_2) : \Big(\frac{v_2h_1 - v_1h_2}{h_1 - h_2}, \frac{v_1 - v_2}{h_1 - h_2}\Big).$$

Thus a necessary condition for optimality of a *differentiated* menu is, $\frac{h_1}{h_2} > 1 + \frac{\lambda_2}{\lambda_1}$. Moreover, price p_2 needs to be non-negative, which implies, $\frac{h_1}{h_2} > \frac{v_1}{v_2}$. Otherwise, $K^* = 1$, i.e. it is optimal to just offer service grade 1. If, on the other hand, $\frac{h_1}{h_2} \leq 1 + \frac{\lambda_2}{\lambda_1}$, from (A.54) we note that the optimal revenue would be $v_2(\lambda_1+\lambda_2)$ which would correspond to $d_1 = d_2 = 0$ and $p_1 = p_2 = v_2$, which is the optimal revenue for the *pooling* menu. Finally, the *pooling menu* would generate higher revenue compared to offering a single service grade catering to customer class 1 and ignoring class 2, if the valuations and arrival rates are such that $\lambda_1 v_1 < (\lambda_1 + \lambda_2)v_2$. ■

Proof of Lemma 3

We are interested in solving problem (1.5), which is equivalent to problem (1.8) with $N = 2$, under fixed arrival rates. Using Lemma 5, we know that the optimal solution to (1.8) is such that $\overline{\lambda}_1 v_1 = \overline{\lambda}_2 v_2$, and $\frac{h_1}{h_2} \to \infty$. Thus, the optimal solution to problem (1.5) is such that

$$R_2^1 = \frac{\lambda_1 v_1}{\lambda_1 v_1 + \lambda_2 v_2} = \frac{\lambda_1 v_1}{\lambda_1 v_1 + \frac{\lambda_2\lambda_1 v_1}{\lambda_1+\lambda_2}} = \frac{\lambda_1 + \lambda_2}{\lambda_1 + 2\lambda_2}.$$

■

Proof of Lemma 4

Let us assume that the WCOM, with N customer classes when the firm offers a single service grade, is such that customer classes i and $i+1$ are pooled into one service grade, and additionally customer class $i-1$ is differentiated, i.e., class $i-1$ joins a service grade that is uniquely different and hence it is not part of the same pool, where $i \geq 2$. We denote the price paid by customer class $i-1$ as p_{i-1} and the price paid by the pooled classes i and $i+1$ as $\hat{p}$ such that $p_{i-1} > \hat{p} > 0$. We are going to prove the statement of the lemma by contradiction. To this end, we will show that there exist incentive compatible prices p_1 and p_2 such that $p_{i-1} > p_1 > \hat{p} > p_2 > 0$, which differentiates customer class i and $i+1$ (i.e. customer class i and $i+1$ join different service grades where customer class i pays p_1 and class $i+1$ pays p_2) as opposed to pooling, and results in a menu with a higher revenue. This creates a contradiction to our assumption that the menu that pools class i and class $i+1$ is the WCOM. Lemma 1 establishes that WCOM belongs to set $\mathbb{M}$. Thus, we can use Proposition 1 to express the revenue-maximizing prices in terms of the valuations and delay sensitivities. Revenue generated from the service grade that pools customer classes i and $i+1$ is given by,

$$r_p = (\lambda_i + \lambda_{i+1})\hat{p} = (\lambda_i + \lambda_{i+1}) v_{i+1} \frac{\frac{h_{i-1}}{h_{i+1}} - \frac{v_{i-1}}{v_{i+1}}}{\frac{h_{i-1}}{h_{i+1}} - 1} = (\lambda_i + \lambda_{i+1}) v_{i+1} \frac{\hat{\delta} - \hat{v}}{\hat{\delta} - 1} \tag{A.55}$$

where $\hat{\delta} = \frac{h_{i-1}}{h_{i+1}}$ and $\hat{v} = \frac{v_{i-1}}{v_{i+1}}$. We are interested in solving problem (1.17). Since the firm offers a single service grade, π_N^1 is given by $\pi_N^1 = \max_{k \in \{1,2,\ldots,N\}} \overline{\lambda}_k v_k$. We can reformulate

this problem as,

$$\begin{aligned} &\inf_{v>0,h>0} z \\ \text{s.t. } z \geq \frac{\overline{\lambda}_1 v_1}{\pi_N^*},\ z \geq \frac{\overline{\lambda}_2 v_2}{\pi_N^*},\ \cdots\ z \geq \frac{\overline{\lambda}_N v_N}{\pi_N^*}&, \\ \Delta_h \geq \prod_{k=1}^{N-1} \delta_k,\ \delta_k \geq 1,\ v_k > v_{k+1} \text{ for all } \ 1 \leq k \leq N-1&, \end{aligned}$$

where $\delta_k = \frac{h_{k-1}}{h_k}$. The Lagrangian for this problem is given by,

$$\begin{aligned} \mathcal{L} = z(1-\alpha_1-\alpha_2-\cdots-\alpha_N) + \frac{\alpha_1\overline{\lambda}_1 v_1 + \alpha_2\overline{\lambda}_2 v_2 + \cdots + \alpha_N\overline{\lambda}_N v_N}{\pi_N^*} \\ + \mu(\prod_{k=1}^{N-1}\delta_k - \Delta_h) + \sum_{k=1}^{N-1}\beta_k(1-\delta_k), \end{aligned}$$

where α_k, β_k and μ are the Lagrangian multipliers. Taking the first order optimality (KKT) conditions we get, $\nabla_z \mathcal{L} = 1-\alpha_1-\alpha_2-\cdots-\alpha_N = 0$, which implies that not all α_k are zero. Furthermore, $\nabla_{v_k}\mathcal{L} = \frac{\pi_N^*\cdot(\alpha_k\overline{\lambda}_k) - \pi_N^1\cdot\frac{\partial\pi_N^*}{\partial v_k}}{(\pi_N^*)^2} = 0$ implies

$$\alpha_k\overline{\lambda}_k = \frac{\pi_N^1}{\pi_N^*}\frac{\partial\pi_N^*}{\partial v_k}. \tag{A.56}$$

We note that the expression for the optimal revenue, π_N^*, does not involve v_i and h_i, which can be seen in the expression for price $\hat{p}$ for this service grade as given by (A.55). This implies $\frac{\partial\pi_N^*}{\partial v_i} = 0$. However, $\frac{\partial\pi_N^*}{\partial v_{i-1}}$ and $\frac{\partial\pi_N^*}{\partial v_{i+1}}$ are clearly non-zero since the prices in the optimal menu change if either v_{i-1} or v_{i+1} is changed. Since dual feasibility condition implies $\alpha_k \geq 0$, from (A.56) we have $\alpha_i = 0, \alpha_{i-1} > 0$ and $\alpha_{i+1} > 0$. By complimentary slackness condition, the constraints corresponding to strictly positive Lagrange multipliers would be strictly binding at optimality which implies $\overline{\lambda}_{i-1}v_{i-1} = \overline{\lambda}_{i+1}v_{i+1}$. Thus, the

optimal values of this problem (worst case valuations and delay sensitivities) are such that,

$$\hat{v} = \frac{v_{i-1}}{v_{i+1}} = \frac{\overline{\lambda}_{i+1}}{\overline{\lambda}_{i-1}}. \tag{A.57}$$

Moreover, since $\alpha_i = 0$, by complementary slackness condition $\overline{\lambda}_i v_i \leq \max_{k\in\{1,2,\ldots,N\}} \overline{\lambda}_k v_k$, i.e., the i^{th} constraint is weakly binding. In order to come up with the new prices, p_1 and p_2, for the differentiated service grades, we will ensure that the prices are incentive compatible and the i^{th} constraint remains weakly binding so that π_N^1 does not change. If we manage to show that π_N^* increases as a result, this will imply that the ratio $\frac{\pi_N^1}{\pi_N^*}$ decreases, proving our claim.

We use Proposition 1 to express the prices p_1 and p_2 in terms of the valuations and the delay sensitivities. Price p_1 is given by $p_1 = \frac{v_i h_{i-1} - v_{i-1} h_i}{h_{i-1} - h_i}$ and price p_2 is given by $p_2 = \frac{v_{i+1} h_i - v_i h_{i+1}}{h_i - h_{i+1}}$. We express v_i as $v_i = (1+\epsilon_v) v_{i+1}$ where $0 < \epsilon_v < \bar{\epsilon}_v$, and $\bar{\epsilon}_v$ ensures that $\overline{\lambda}_i v_i < \max_{k\in\{1,2,\ldots,N\}} \overline{\lambda}_k v_k$. Moreover, we express h_i as $h_i = (1+\epsilon_h) h_{i+1}$ where $0 < \epsilon_h < \bar{\epsilon}_h$ and $\bar{\epsilon}_h = \hat{\delta} - 1$ ensures that $h_i < h_{i-1}$. Our goal would be to show that, leaving all other parameters unchanged, there exist feasible values of ϵ_v and ϵ_h which will result in the prices p_1 and p_2 such that $(\lambda_i + \lambda_{i+1})\hat{p} < \lambda_i p_1 + \lambda_{i+1} p_2$. The revenue generated from differentiating customer class i and $i+1$ is given by,

$$\begin{aligned} r_s = \lambda_i p_1 + \lambda_{i+1} p_2 \; &= \lambda_i \frac{v_i h_{i-1} - v_{i-1} h_i}{h_{i-1} - h_i} + \lambda_{i+1} \frac{v_{i+1} h_i - v_i h_{i+1}}{h_i - h_{i+1}} \\ &= \lambda_i v_i \frac{\frac{h_{i-1}}{h_i} - \frac{v_{i-1}}{v_i}}{\frac{h_{i-1}}{h_i} - 1} + \lambda_{i+1} v_{i+1} \frac{\frac{h_i}{h_{i+1}} - \frac{v_i}{v_{i+1}}}{\frac{h_i}{h_{i+1}} - 1}, \end{aligned}$$

which is equivalent to,

$$r_s = \lambda_i v_{i+1}\Big(\frac{(1+\epsilon_v)\hat{\delta} - \hat{v}(1+\epsilon_h)}{\hat{\delta} - (1+\epsilon_h)}\Big) + \lambda_{i+1} v_{i+1}\Big(1 - \frac{\epsilon_v}{\epsilon_h}\Big). \tag{A.58}$$

In order to maintain incentive compatibility, the new service grade offering price $p_1 > \hat{p}$ would offer a lower delay as compared to the pooling service grade with price $\hat{p}$. Similarly, the new service grade with price $p_2 < \hat{p}$ would offer higher delay. The constraint on the new price $p_2 < \hat{p}$ implies, $\frac{v_{i+1}h_i - v_i h_{i+1}}{h_i - h_{i+1}} < \frac{v_{i+1}h_{i-1} - v_{i-1}h_{i+1}}{h_{i-1} - h_{i+1}}$. Using the previously defined expressions $v_i = (1+\epsilon_v)v_{i+1}, h_i = (1+\epsilon_h)h_{i+1}, \frac{h_{i-1}}{h_{i+1}} = \hat{\delta}$ and $\frac{v_{i-1}}{v_{i+1}} = \hat{v}$, we have $v_{i+1}\big(1-\frac{\epsilon_v}{\epsilon_h}\big) < v_{i+1}\frac{\hat{\delta}-\hat{v}}{\hat{\delta}-1}$ which is equivalent to

$$\frac{\hat{v}-1}{\hat{\delta}-1} < \frac{\epsilon_v}{\epsilon_h}.$$

Let us denote $\frac{\epsilon_v}{\epsilon_h} - \frac{\hat{v}-1}{\hat{\delta}-1} = \xi$. We will choose $\xi \to 0$ such that $p_2 \to \hat{p}$. From (A.55) and (A.58), we have,

$$\begin{aligned}
r_s - r_p &= \lambda_i p_1 + \lambda_{i+1} p_2 - (\lambda_i + \lambda_{i+1})\hat{p} \\
&= \lambda_i v_{i+1}\Big(\frac{(1+\epsilon_v)\hat{\delta} - \hat{v}(1+\epsilon_h)}{\hat{\delta} - (1+\epsilon_h)}\Big) + \lambda_{i+1} v_{i+1}\Big(1 - \frac{\epsilon_v}{\epsilon_h}\Big) - (\lambda_i + \lambda_{i+1}) v_{i+1}\frac{\hat{\delta} - \hat{v}}{\hat{\delta} - 1} \\
&= v_{i+1}\Big(\lambda_i \frac{\hat{\delta} - \hat{v}}{\hat{\delta} - 1 - \epsilon_h} - \lambda_i \frac{\hat{\delta} - \hat{v}}{\hat{\delta} - 1} + \overbrace{\lambda_i \frac{\epsilon_v\hat{\delta} - \epsilon_h\hat{v}}{\hat{\delta} - 1 - \epsilon_h}}^{Replacing \frac{\epsilon_v}{\epsilon_h} = \xi + \frac{\hat{v}-1}{\hat{\delta}-1}} \\
&\quad + \underbrace{\lambda_{i+1} - \lambda_{i+1}\frac{\hat{v}-1}{\hat{\delta}-1} - \lambda_{i+1}\frac{\hat{\delta}-\hat{v}}{\hat{\delta}-1}}_{=0} - \lambda_{i+1}\xi\Big) \\
&= v_{i+1}\Big(\lambda_i \epsilon_h \frac{\hat{\delta} - \hat{v}}{(\hat{\delta}-1)(\hat{\delta}-1-\epsilon_h)} + \epsilon_h \lambda_i \frac{\hat{\delta}(\xi + \frac{\hat{v}-1}{\hat{\delta}-1}) - \hat{v}}{\hat{\delta} - 1 - \epsilon_h} - \lambda_{i+1}\xi\Big)
\end{aligned}$$

$$
\begin{aligned}
&= v_{i+1}\Big(\lambda_i\epsilon_h\xi\frac{\hat{\delta}}{\hat{\delta}-1-\epsilon_h}-\lambda_{i+1}\xi\Big)\\
&= v_{i+1}\Big(\xi\frac{\epsilon_h(\lambda_i\hat{\delta}+\lambda_{i+1})-\lambda_{i+1}(\hat{\delta}-1)}{\hat{\delta}-1-\epsilon_h}\Big). \qquad \text{(A.59)}
\end{aligned}
$$

The denominator in (A.59) is positive since $\epsilon_h < \bar{\epsilon}_h = \hat{\delta}-1$. The numerator is positive for all $\epsilon_h > \frac{\lambda_{i+1}(\hat{\delta}-1)}{\lambda_i\hat{\delta}+\lambda_{i+1}} = \underline{\epsilon}_h$. We note that, for $\lambda_i > 0$, we have feasible ϵ_h such that $\bar{\epsilon}_h > \epsilon_h > \underline{\epsilon}_h$ which would result in $r_s - r_p > 0$. Now, we also need to show that there exist feasible ϵ_v such that $\bar{\epsilon}_v > \epsilon_v > \underline{\epsilon}_v$ which would be sufficient for $r_s - r_p$ to be positive. To this end, from our definition of $\xi = \frac{\epsilon_v}{\epsilon_h} - \frac{\hat{v}-1}{\hat{\delta}-1}$, we can write $\epsilon_v = \Big(\xi + \frac{\hat{v}-1}{\hat{\delta}-1}\Big)\epsilon_h$. Plugging in $\underline{\epsilon}_h = \frac{\lambda_{i+1}(\hat{\delta}-1)}{\lambda_i\hat{\delta}+\lambda_{i+1}}$ for ϵ_h in $\epsilon_v = \Big(\xi+\frac{\hat{v}-1}{\hat{\delta}-1}\Big)\epsilon_h$, we get $\epsilon_v > \Big(\xi\frac{\lambda_{i+1}(\hat{\delta}-1)}{\lambda_i\hat{\delta}+\lambda_{i+1}}+\frac{\lambda_{i+1}(\hat{v}-1)}{\lambda_i\hat{\delta}+\lambda_{i+1}}\Big)$. This provides us with the expression for $\underline{\epsilon}_v$. Thus, we have,

$$
\underline{\epsilon}_v = \xi\frac{\lambda_{i+1}(\hat{\delta}-1)}{\lambda_i\hat{\delta}+\lambda_{i+1}} + \frac{\lambda_{i+1}(\hat{v}-1)}{\lambda_i\hat{\delta}+\lambda_{i+1}}. \qquad \text{(A.60)}
$$

We have stated earlier that $\bar{\epsilon}_v$ ensures that $\overline{\lambda}_i v_i < \max_{k\in\{1,2,\ldots,N\}}\overline{\lambda}_k v_k$ holds true such that π_N^1 remains unchanged while we set v_i appropriately. A sufficient condition for $\overline{\lambda}_i v_i < \max_{k\in\{1,2,\ldots,N\}}\overline{\lambda}_k v_k$ would be $\overline{\lambda}_i v_i < \overline{\lambda}_{i+1} v_{i+1}$ since the latter implies the former. Plugging in $v_i = (1+\epsilon_v)v_{i+1}$ in $\overline{\lambda}_i v_i < \overline{\lambda}_{i+1} v_{i+1}$ implies $\epsilon_v < \frac{\lambda_{i+1}}{\overline{\lambda}_i}$. Thus, we define,

$$
\bar{\epsilon}_v = \frac{\lambda_{i+1}}{\overline{\lambda}_i}. \qquad \text{(A.61)}
$$

All we need to show is that $\bar{\epsilon}_v - \underline{\epsilon}_v > 0$, which will imply the existence of a feasible ϵ_v. From (A.60) and (A.61) we have the following expression: $\bar{\epsilon}_v - \underline{\epsilon}_v = \frac{\lambda_{i+1}}{\overline{\lambda}_i} - \frac{\lambda_{i+1}(\hat{v}-1)}{\lambda_i\hat{\delta}+\lambda_{i+1}} - \xi\frac{\lambda_{i+1}(\hat{\delta}-1)}{\lambda_i\hat{\delta}+\lambda_{i+1}}$. From (A.57) we have, $\hat{v} = \frac{v_{i-1}}{v_{i+1}} = \frac{\overline{\lambda}_{i+1}}{\overline{\lambda}_{i-1}}$. Replacing this in the above expression, we can

rewrite it as,

$$\begin{aligned}
\bar{\epsilon}_v - \underline{\epsilon}_v &= \frac{\lambda_{i+1}}{\bar{\lambda}_i} - \frac{\lambda_{i+1}(\lambda_i + \lambda_{i+1})}{\bar{\lambda}_{i-1}(\lambda_i\hat{\delta} + \lambda_{i+1})} - \xi\frac{\lambda_{i+1}(\hat{\delta}-1)}{\lambda_i\hat{\delta} + \lambda_{i+1}} \\
&= \frac{\lambda_i\lambda_{i+1}\bar{\lambda}_{i-1}\left(\hat{\delta} + \frac{\lambda_{i+1}}{\lambda_i} - \frac{\bar{\lambda}_i(\lambda_i+\lambda_{i+1})}{\lambda_i\bar{\lambda}_{i-1}}\right)}{\bar{\lambda}_i\bar{\lambda}_{i-1}(\lambda_i\hat{\delta} + \lambda_{i+1})} - \xi\frac{\lambda_{i+1}(\hat{\delta}-1)}{\lambda_i\hat{\delta} + \lambda_{i+1}} \\
&= \underbrace{\frac{\lambda_i\lambda_{i+1}\bar{\lambda}_{i-1}\left(\hat{\delta} - \frac{\bar{\lambda}_{i+1}}{\bar{\lambda}_{i-1}}\right)}{\bar{\lambda}_i\bar{\lambda}_{i-1}(\lambda_i\hat{\delta} + \lambda_{i+1})}}_{>0} - \xi\frac{\lambda_{i+1}(\hat{\delta}-1)}{\lambda_i\hat{\delta} + \lambda_{i+1}}.
\end{aligned}$$

We know from Proposition 1 that if classes $i-1$ and $i+1$ are differentiated then $\frac{h_{i-1}}{h_{i+1}} = \hat{\delta} > \frac{\bar{\lambda}_{i+1}}{\bar{\lambda}_{i-1}}$ holds true. This implies that the first term is strictly positive. We can choose sufficiently small $\xi \to 0$ to make the above expression for $\bar{\epsilon}_v - \underline{\epsilon}_v$ strictly positive. This suggests that we can choose prices p_1 and p_2 by choosing ϵ_h and ϵ_v such that $\bar{\epsilon}_v > \epsilon_v > \underline{\epsilon}_v$ and $\bar{\epsilon}_h > \epsilon_h > \underline{\epsilon}_h$, holding all other parameters fixed, resulting in an increase in the optimal revenue π_N^* which keeping π_N^1 unchanged. This in turn results in a lower revenue ratio. Hence, our assumption that the WCOM pools customer classes i and i+1 and differentiates class $i-1$, leads to a contradiction. This implies that the customer segmentation induced by the WCOM is such that all customer classes that join distinctly separate service grades form a continuous block in which no two customer classes are pooled into the same service grade. ■

Lemma 5. *The optimal solution to problem (1.8) is such that* $\frac{h_k}{h_{k+1}} \to \infty$ *and* $\overline{\lambda}_k v_k = \overline{\lambda}_{k+1} v_{k+1}$ *for all* $k = 1, 2, \ldots, N-1$, *where* $\overline{\lambda}_k = \sum_{i=1}^{k} \lambda_i$.

Proof. The optimization problem under consideration is as follows:

$$R_N^1 = \inf_{\lambda>0,v>0,h>0} \left\{ \frac{\max_{1\le k\le N}\{\overline{\lambda}_k v_k\}}{\lambda_1 v_1 + \sum_{k=2}^{N} \lambda_k \big(v_k - h_k\big(\frac{v_{k-1}-v_k}{h_{k-1}-h_k}\big)\big)} \right\}$$

$$\text{s.t. } v_k > v_{k+1}, \tag{A.62}$$

$$\frac{h_k}{h_{k+1}} > \frac{v_k}{v_{k+1}}, \text{ for all } k = 1, 2, \ldots, N-1.$$

First, let's assume that the worst case delay sensitivities, under the optimal solution of (A.62), are such that $\frac{h_k}{h_{k+1}}$ is finite for any $k = 1, 2, \ldots, N-1$. Then for any arbitrarily small $\epsilon > 0$, $\frac{h_k}{h_{k+1}} + \epsilon$ would increase the value of the denominator of the objective function in (A.62), $\lambda_1 v_1 + \sum_{k=2}^{N} \lambda_k \big(v_k - h_k\big(\frac{v_{k-1}-v_k}{h_{k-1}-h_k}\big)\big)$, and thus lowers the value of the objective function. This creates a contradiction since (A.62) is a minimization problem. Hence, the worst case delay sensitivities are such that $\frac{h_k}{h_{k+1}} \to \infty$ for all $k = 1, 2, \ldots, N-1$. Now, we can rewrite (A.62) as

$$R_N^1 = \inf_{\lambda>0,v>0} \left\{ \frac{\max_{1\le k\le N}\{\overline{\lambda}_k v_k\}}{\sum_{k=1}^{N} \lambda_k v_k} \right\}.$$

We can reformulate the above optimization problem as

$$\inf_{\lambda>0,v>0} z$$

$$\text{s.t. } z \ge \frac{\overline{\lambda}_k v_k}{\lambda_1 v_1 + \lambda_2 v_2 + \cdots + \lambda_N v_N}, \quad \text{for all } k = 1, 2, \ldots, N,$$

$$v_k > v_{k+1}, \quad \text{for all } k = 1, 2, \ldots, N-1.$$

The Lagrangian for the above problem is

$$\mathcal{L} = z - \mu_1\Big(z - \frac{\overline{\lambda}_1 v_1}{\lambda_1 v_1 + \lambda_2 v_2 + \cdots + \lambda_N v_N}\Big) - \mu_2\Big(z - \frac{\overline{\lambda}_2 v_2}{\lambda_1 v_1 + \lambda_2 v_2 + \cdots + \lambda_N v_N}\Big) - \cdots$$

$$- \mu_N \Big(z - \frac{\overline{\lambda}_N v_N}{\lambda_1 v_1 + \lambda_2 v_2 + \cdots + \lambda_N v_N} \Big)$$

$$= z(1 - \mu_1 - \mu_2 - \cdots - \mu_N) + \frac{\mu_1 \overline{\lambda}_1 v_1 + \mu_2 \overline{\lambda}_2 v_2 + \cdots + \mu_N \overline{\lambda}_N v_N}{\lambda_1 v_1 + \lambda_2 v_2 + \cdots + \lambda_N v_N}$$

where $\mu_1, \mu_2, \ldots, \mu_N$ are the Lagrange multipliers. Taking the first order optimality (KKT) conditions, $\nabla_z \mathcal{L} = 0$ implies $\mu_1 + \mu_2 + \cdots + \mu_N = 1$, and $\nabla_{v_k} \mathcal{L} = 0$ implies

$$\mu_k = \frac{\mu_1 \overline{\lambda}_1 v_1 + \mu_2 \overline{\lambda}_2 v_2 + \cdots + \mu_N \overline{\lambda}_N v_N}{\lambda_1 v_1 + \lambda_2 v_2 + \cdots + \lambda_N v_N} . \frac{\lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_k} . \tag{A.63}$$

From the dual feasibility conditions, $\mu_k \geq 0$ for all k. From the first optimality condition, $\mu_1 + \mu_2 + \cdots + \mu_N = 1$ implies that not all μ_k can be zero. Since the right hand side of equation (A.63) is strictly positive, μ_k is strictly positive as well, i.e., $\mu_k > 0$ for all k. Using complementary slackness condition, all constraints with their corresponding Lagrange multipliers are binding at optimality. Therefore, $\overline{\lambda}_1 v_1 = \overline{\lambda}_2 v_2 = \cdots = \overline{\lambda}_N v_N$. ■

Lemma 6. *The optimal solution to problem (A.10) under constraints (A.11), (A.12) and (A.13), is such that $c_k \to \infty$ for all $k = 1, 2, \ldots, \infty$.*

Proof. First, looking at constraint (A.12) we note that, c_k cannot be finite for all $k = 1, 2, \ldots, \infty$. The reason for this is that the left-hand-side of inequality (A.12) is an infinite product and the right-hand-side is a finite number Δ_v. Thus, under the optimal solution, c_k can be finite for only finitely many k. Let us assume that, for any $\Delta_h > 1$ and $\Delta_v > 1$, the optimal solution to (A.10) is such that c_k is finite for finitely many k. To this end, we define a finite set of indices, $\mathcal{K}$, such that, under the optimal solution to (A.10), $c_k \leq C$ for all $k \in \mathcal{K}$, and $c_k \to \infty$ for all $k \notin \mathcal{K}$, for some arbitrary positive

constant $0 < C < \infty$. Using this, let's rewrite the optimization problem as,

$$\inf_{\{\delta_k, c_k | k \in \mathcal{K}\}} \Big\{ \sum_{k \in \mathcal{K}} \Big(\frac{1}{c_k(c_k+1)(\delta_k - 1)} - \frac{1}{1+c_k} \Big) + f(\{\delta_k, c_k | k \in \mathcal{K}\}) \Big\}$$

$$\text{s.t.} \prod_{k \in \mathcal{K}} \delta_k \leq \Delta_h,$$

$$\prod_{k \in \mathcal{K}} \Big(1 + \frac{1}{c_k} \Big) \leq \Delta_v,$$

$$\delta_k > 1 + \frac{1}{c_k}, c_k > 0 \text{ for all } k \in \mathcal{K},$$

where $f(\{\delta_k, c_k | k \in \mathcal{K}\})$ is given by,

$$f(\{\delta_k, c_k | k \in \mathcal{K}\}) = \inf_{\{\delta_k, c_k | k \notin \mathcal{K}\}} \sum_{k \notin \mathcal{K}} \Big(\frac{1}{c_k(c_k+1)(\delta_k - 1)} - \frac{1}{1+c_k} \Big)$$

$$\text{s.t.} \prod_{k \notin \mathcal{K}} \delta_k \leq \tilde{\Delta}_h,$$

$$\prod_{k \notin \mathcal{K}} \Big(1 + \frac{1}{c_k} \Big) \leq \tilde{\Delta}_v,$$

$$\delta_k > 1 + \frac{1}{c_k}, c_k > 0 \text{ for all } k \notin \mathcal{K}$$

where $\tilde{\Delta}_h = \dfrac{\Delta_h}{\prod_{k \in \mathcal{K}} \delta_k}$ and $\tilde{\Delta}_v = \dfrac{\Delta_v}{\prod_{k \in \mathcal{K}} \left(1 + \frac{1}{c_k} \right)}$.

Now, we can rename the decision variables in optimization problem $f(\{\delta_k, c_k | k \in \mathcal{K}\})$ and

rewrite it as,

$$
\begin{aligned}
f(\{\delta_k, c_k | k \in \mathcal{K}\}) = \inf_{\tilde{\delta}_k, \tilde{c}_k} & \sum_{k=1}^{\infty} \Big(\frac{1}{\tilde{c}_k(\tilde{c}_k+1)(\tilde{\delta}_k - 1)} - \frac{1}{1+\tilde{c}_k} \Big) \\
\text{s.t.} \ & \prod_{k=1}^{\infty} \tilde{\delta}_k \leq \tilde{\Delta}_h, \\
& \prod_{k=1}^{\infty} \Big(1 + \frac{1}{\tilde{c}_k}\Big) \leq \tilde{\Delta}_v, \\
& \tilde{\delta}_k > 1 + \frac{1}{\tilde{c}_k}, \tilde{c}_k > 0 \text{ for all } k = 1, 2, \ldots, \infty.
\end{aligned} \tag{A.64}
$$

Problem (A.64) is equivalent to the original optimization problem (A.10) with bounds $\tilde{\Delta}_h$ and $\tilde{\Delta}_v$. Hence, by our assumption on the optimal solution to (A.10), $\tilde{c}_k$ should be finite for finitely many k. However, by definition of set $\mathcal{K}$, the optimal solution to problem $f(\{\delta_k, c_k | k \in \mathcal{K}\})$ is such that $c_k \to \infty$ for all $k \notin \mathcal{K}$. This creates a contradiction. Hence, the optimal solution to (A.10) is such that $c_k \to \infty$ for all $k = 1, 2, \ldots, \infty$. ■

Lemma 7. *The optimal solution to the optimization problem (1.21) is achieved as* $i^* \to \infty$.

Proof. We will show that $\frac{1}{4}\log(\Delta_h) - \sum_{k=i}^{j-1} \big(\frac{1}{1+c_k} - \frac{1}{c_k(c_k+1)(\delta_k - 1)} \big) > 0$ if i is finite. Thus, we consider the following function:

$$
G_{ij}(\delta_i, \delta_{i+1}, \ldots, \delta_{j-1}) = \frac{1}{4}\log(\Delta_h) - \sum_{k=i}^{j-1} \left(\frac{1}{1+c_k} - \frac{1}{c_k(c_k+1)(\delta_k - 1)} \right)
$$

where $\prod_{k=i}^{j-1} \delta_k = \Delta_h$ from (A.4). Using this, we replace $\log(\Delta_h)$ in G_{ij} with $\sum_{k=i}^{j-1} \log(\delta_k)$. Thus we have,

$$G_{ij}(\delta_i, \delta_{i+1}, \ldots, \delta_{j-1}) = \sum_{k=i}^{j-1} \left(\frac{1}{4} \log(\delta_k) - \frac{1}{1+c_k} + \frac{1}{c_k(c_k+1)(\delta_k - 1)} \right).$$

We denote the k^{th} term in the above summation as

$$G_{ij}^k(\delta_k) = \big(\frac{1}{4} \log(\delta_k) - \frac{1}{1+c_k} + \frac{1}{c_k(c_k+1)(\delta_k - 1)} \big).$$

Taking the price condition, $\delta_k > 1 + \frac{1}{c_k}$, into consideration the function $G_{ij}^k(\delta_k)$ has a unique global minimum. Taking the first order stationary condition with respect to δ_k, the stationary point is given by the following equation:

$$\frac{1}{4\delta_k} - \frac{1}{c_k(c_k+1)(\delta_k - 1)^2} = 0.$$

Solving for $\delta_k > 1 + \frac{1}{c_k}$, we get,

$$\delta_k^* = 1 + \frac{2 + 2(c_k^2 + c_k + 1)^{\frac{1}{2}}}{c_k(c_k+1)}.$$

Replacing the value of δ_k in $G_{ij}^k(\delta_k)$ with δ_k^*, we have,

$$G_{ij}^k(\delta_k^*) = \frac{1}{4} \log \left(1 + \frac{2(c_k^2 + c_k + 1)^{\frac{1}{2}} + 2}{c_k^2 + c_k} \right) - \frac{1}{c_k + 1} + \frac{1}{2(c_k^2 + c_k + 1)^{\frac{1}{2}} + 2}.$$

For any $c_k > 0$, $G_{ij}^k(\delta_k^*)$ has a strictly positive value which strictly decreases as c_k increases, and asymptotically goes to zero as $c_k \to \infty$. Thus, $G_{ij}(\delta_i^*, \delta_{i+1}^*, \ldots, \delta_{j-1}^*) \to 0$ if $G_{ij}^k(\delta_k^*) \to 0$ for all $k = i^*, i^* + 1, \ldots, j^* - 1$. Since $\{c_k\}_{k=1}^{\infty}$ is a diverging sequence, it implies $G_{ij}(\delta_i^*, \delta_{i+1}^*, \ldots, \delta_{j-1}^*) \to 0$ as $i^* \to \infty$. ■

Proof of Corollary 1

Theorem 2 presents the solution to problem (1.21), and the WCRR is given by, $R^1_\infty(\Delta_h) = \frac{1}{1+\frac{1}{4}\log(\Delta_h)}$. Thus, for any given sequence of arrival rates, $\{\lambda_k\}_{k=1}^\infty$, the value of $R^1_\infty(\Delta_h)$ does not depend on λ_k or c_k for any k. ■

APPENDIX B

Technical Analysis and Proofs of Results in Chapter 2

B.1. Expected Wait for Geometric Service Slot Distribution

We use $\mathbb{E}W(k)$ to denote the expected number of time periods an app users or a non-app user has to wait in store for the completion of k orders, including her own. We will use the shorthand W_k to denote $\mathbb{E}W(k)$. Here, we derive the analytical expression for W_k when $S_t \sim$ Geometric (p), where S_t denotes the number of service slots generated in period t. Since S_t is i.i.d., W_k will satisfy the following recursive equation:

$$W_k = \mathbb{P}(S_t \geq k) \cdot 0 + \sum_{i=0}^{k-1} \mathbb{P}(S_t = i) \cdot (1 + W_{k-i}) \tag{B.1}$$

The expected value of S_t is $\mu = \frac{1-p}{p}$, where $\mathbb{P}(S_t = i) = p(1-p)^i$. We will prove that $W_k = \frac{k}{\mu}$ by induction. First, using (B.1), we have $W_1 = \mathbb{P}(S_t = 0) \cdot (1 + W_1)$, which implies $W_1 = \frac{1}{\mu}$. Next, by replacing W_{k-i} by $\frac{k-i}{\mu}$ in (B.1) we can rewrite the RHS as,

$$\begin{aligned} \sum_{i=0}^{k-1} p(1-p)^i \cdot \left(1 + \frac{k-i}{\mu}\right) &= \sum_{i=0}^{k-1} p(1-p)^i \cdot \left(1 + \frac{k-i}{1-p}p\right) \\ &= \left(p + \frac{kp^2}{1-p}\right) \cdot \sum_{i=0}^{k-1}(1-p)^i - p^2 \cdot \sum_{i=1}^{k-1} i(1-p)^{i-1} \end{aligned} \tag{B.2}$$

We use the following two formulae to simplify (B.2):

$$\sum_{i=0}^{k-1} r^i = \frac{1-r^k}{1-r}, \qquad \sum_{i=1}^{k-1} ir^{i-1} = \frac{1-r^k}{(1-r)^2} - \frac{kr^{k-1}}{1-r}. \tag{B.3}$$

Using (B.3), we can rewrite (B.2) as,

$$\begin{aligned}
&\left(p+\frac{kp^2}{1-p}\right)\cdot\sum_{i=0}^{k-1}(1-p)^i - p^2\cdot\sum_{i=1}^{k-1} i(1-p)^{i-1}\\
=&\left(p+\frac{kp^2}{1-p}\right)\cdot\left(\frac{1-(1-p)^k}{p}\right) - p^2\cdot\left(\frac{1}{p^2}(1-(1-p)^k) - \frac{k}{p}(1-p)^{k-1}\right)\\
=&(1-(1-p)^k) + \frac{kp}{1-p} - kp(1-p)^{k-1} - (1-(1-p)^k) + kp(1-p)^{k-1}\\
=&\frac{kp}{1-p} = \frac{k}{\mu}. \qquad \text{(B.4)}
\end{aligned}$$

Thus, from (B.4), we have the RHS in (B.1), and we get $W_k = \frac{k}{\mu}$. ■

B.2. Steady State Analysis of Markov Chain

To compute firm's throughput and the average consumer surplus, we will construct a Markov Chain using the events described in §2.2.2. Recall that in every time period t, four events take place, Event 1 (non-app users arrive at the store), Event 2 (app users from previous period arrives at the store), Event 3 (a new app user arrives at the market) and Event 4 (services take place), as defined in §2.2.2. Length of the queue in period t before Event $i+1$ is denoted by $x_i(t)$ where $i = 0, 1, 2, 3$. Additionally, we use $u_i(t)$ to track whether or not there is an undecided app user in period t; presence of an undecided app user in period t before Event $i+1$ is denoted by $u_i(t)$ where $i = 0, 1, 2, 3$. In particular, at the beginning of time period t, if there is an undecided app user from previous period $t-1$ traveling to the store, we denote it by $u_0(t) = 1$. This undecided app user decides whether to order or not after observing the queue at the store in period t. Note that, $u_0(t) = 0$ denotes that the app user who arrived at the market in previous period $t-1$ ordered online, in which case, this app user travels to the store for pick-up. Next, in

Event 1 non-app users arrive, which does not alter the status of the undecided app user. Therefore, $u_1(t) = u_0(t)$. In Event 2, the undecided app user decides whether to join or balk after observing the queue. Hence, after Event 2, the app user's join/balk decision is resolved, which is denoted by $u_2(t) = 0$ for all t. Note that right after Event 2 and before Event 3, there is never an undecided app user in the system. Finally, in Event 3, a new app user arrives at the market in period t. If this new app user orders online, we denote it by $u_3(t) = 0$. Otherwise if this app user is undecided and defers his joining decision until he arrives at the store in the next time period, we denote it by $u_3(t) = 1$. Thus, we note that $u_i(t)$ can only change its value twice; $u_2(t) = 0$ by definition and $u_3(t)$ reflects the decision made by the app user arriving at the market. We henceforth suppress the time argument in $x_i(t)$ and $u_i(t)$ and use x_i and u_i to denote steady state. Finally, we denote the state of the system by (x_i, u_i), where the value of i could be chosen as $0, 1, 2$ or 3. Note that (x_i, u_i) is a Markov Chain. Due to the sequence of four events, as defined in §2.2.2, we decompose every state transition into four intermediate transitions. We begin by presenting the transition probability matrices associated with the state transitions.

Arrival of Non-App Users. The intermediate transition from (x_0, u_0) to (x_1, u_1) takes place as a result of arrivals of non-app users at the store. The probability of k arrivals of non-app users at the store in period t is denoted by $\mathbb{P}(A_t = k) = a_k$. For brevity, we adopt the notation $\sum_{i=k}^{\infty} a_i = \overline{a}_k$. We know from (2.5) that the non-app users follow a threshold-based joining policy, upon arrival at the store and observing the queue length, x_0. Non-app users join only if the number of orders in the system, x_0, is strictly less than τ_n. The following matrix, $\mathcal{M}_0$, holds the transition probabilities, where the row indices

correspond to x_0, the column indices correspond to x_1. Matrix $\mathcal{M}_0$ is given by,

$$\mathcal{M}_0 = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ \vdots \\ \tau_n-1 \\ \tau_n \\ \tau_n+1 \\ \vdots \\ M \end{array} \begin{array}{c} \begin{array}{ccccccccc} 0 & 1 & 2 & \cdots & \tau_n-1 & \tau_n & \tau_n+1 & \cdots & M \end{array} \\ \begin{bmatrix} a_0 & a_1 & a_2 & \cdots & a_{\tau_n-1} & \overline{a}_{\tau_n} & 0 & \cdots & 0 \\ 0 & a_0 & a_1 & \cdots & a_{\tau_n-2} & \overline{a}_{\tau_n-1} & 0 & \cdots & 0 \\ 0 & 0 & a_0 & \cdots & a_{\tau_n-3} & \overline{a}_{\tau_n-2} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & a_0 & \overline{a}_1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 1 \end{bmatrix} \end{array}.$$

where M is some arbitrarily large integer. For the model with information, we set $M = \max(\tau_n, \tau_s, \tau_u+1)$. For the model without information we set the value of M large enough such that the probability of the queue length being M or larger is arbitrariliy small.

We note that an intermediate transition from $(x_0, 0)$ can only lead to $(x_1, 0)$ since $u_0 = u_1$. Similarly, a transition from $(x_0, 1)$ can only lead to $(x_1, 1)$. Thus, the transition probability matrix corresponding to intermediate transitions from (x_0, u_0) to (x_1, u_1) is given by,

$$\mathcal{P}_0 = \begin{array}{c} \\ (x_0,0) \\ (x_0,1) \end{array} \begin{array}{c} \begin{array}{cc} (x_1,0) & (x_1,1) \end{array} \\ \begin{bmatrix} \mathcal{M}_0 & \mathbf{0} \\ \mathbf{0} & \mathcal{M}_0 \end{bmatrix} \end{array},$$

where $\mathbf{0}$ in the matrix $\mathcal{P}_0$ denotes a matrix containing zeros, having same dimension as that of $\mathcal{M}_0$.

Arrival of App User at Store. The intermediate transition from (x_1, u_1) to (x_2, u_2), takes place as a result of Event 2. As already mentioned, if $u_1 = 0$, then the app user chose the online option and is at the store for pick-up. Thus, the queue length remains unchanged, i.e. $(x_1, 0)$ transitions to $(x_2, 0)$ where $x_2 = x_1$. On the other hand, if $u_1 = 1$, then the app user chose the offline option and decides whether to order or balk in Event 2. We have established in (2.6) that an app user at the store follows a Naor-type threshold, τ_s, to decide whether or not to join. Thus, $(x_1, 1)$ transitions to state $(x_2, 0)$, where the transition matrix $\mathcal{M}_1$, is given by,

$$\mathcal{M}_1 = \begin{array}{c} \\ 0 \\ 1 \\ \vdots \\ \tau_s-1 \\ \tau_s \\ \tau_s+1 \\ \vdots \\ M \end{array} \begin{array}{c} \begin{array}{cccccccc} 0 & 1 & 2 & \cdots & \tau_s & \tau_s+1 & \cdots & M \end{array} \\ \left[\begin{array}{cccccccc} 0 & 1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & 0 & \cdots & 1 \end{array}\right] \end{array}.$$

The transition probability matrix corresponding to intermediate transitions from (x_1, u_1) to (x_2, u_2) is given by,

$$\mathcal{P}_1 = \begin{array}{c} \\ (x_1,0) \\ (x_1,1) \end{array} \begin{array}{c} \begin{array}{cc} (x_2,0) & (x_2,1) \end{array} \\ \begin{bmatrix} \mathcal{I} & \mathbf{0} \\ \mathcal{M}_1 & \mathbf{0} \end{bmatrix} \end{array},$$

where $\mathcal{I}$ denotes the identity matrix, and $\mathbf{0}$ denotes a matrix containing zeros, both $\mathcal{I}$ and $\mathbf{0}$ having same dimension as that of $\mathcal{M}_1$.

Arrival of App User at the Market. The intermediate transition from (x_2, u_2) to (x_3, u_3) occurs as a result of the arrival of a new app user at the market. The strategy adopted by this app user depends on whether or not the current queue length, x_2, is revealed to her. In Proposition 4, we have established that when queue information is revealed, there are two thresholds, τ_l and τ_u, such that the app user orders online if the queue length, x_2, satisfies $\tau_l \leq x_2 \leq \tau_u$, and chooses the offline option otherwise. Note that depending on the system parameters, τ_l could be zero. Thus, $(x_2, 0)$ transitions to $(x_3, 0)$ when $\tau_l \leq x_2 \leq \tau_u$, whereas, transitions happen to $(x_3, 1)$ otherwise. Note that, we always have $u_2 = 0$, as already mentioned. Also, note that u_3 reflects the online/offline decision made by the new app user arriving at the market. The probability matrix for the intermediate transitions from (x_2, u_2) to (x_3, u_3) is given by,

$$\mathcal{P}_2 = \begin{array}{c} \\ (x_2,0) \\ (x_2,1) \end{array} \begin{array}{c} \begin{array}{cc} (x_3,0) & (x_3,1) \end{array} \\ \begin{bmatrix} \mathcal{M}_{21} & \mathcal{M}_{22} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \end{array},$$

where $\mathcal{M}_{21}$ and $\mathcal{M}_{22}$ are given by,

$$\mathcal{M}_{21} = \begin{array}{c} \\ 0 \\ \vdots \\ \tau_l{-}1 \\ \tau_l \\ \vdots \\ \tau_u \\ \tau_u{+}1 \\ \vdots \\ M \end{array} \begin{array}{c} \begin{array}{cccccccccc} 0 & \cdots & \tau_l & \tau_l{+}1 & \cdots & \tau_u{+}1 & \tau_u{+}2 & \cdots & M \end{array} \\ \begin{bmatrix} 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 1 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \end{bmatrix} \end{array},$$

$$\mathcal{M}_{22} = \begin{array}{c} \\ 0 \\ \vdots \\ \tau_l{-}1 \\ \tau_l \\ \vdots \\ \tau_u \\ \tau_u{+}1 \\ \vdots \\ M \end{array} \begin{array}{c} \begin{array}{cccccccccc} 0 & \cdots & \tau_l{-}1 & \tau_l & \cdots & \tau_u & \tau_u{+}1 & \cdots & M \end{array} \\ \begin{bmatrix} 1 & \cdots & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 0 & \cdots & 0 & 1 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & \cdots & 1 \end{bmatrix} \end{array}.$$

Again, $\mathbf{0}$ in the matrix $\mathcal{P}_2$, denotes a matrix containing zeros, having the same dimension as that of $\mathcal{M}_{21}$ and $\mathcal{M}_{22}$.

Now, if queue-length information is withheld from the app user who arrives at the market, we have established in Theorem 4 that her joining strategy can be denoted by θ, which represents the probability with which she orders online. Accordingly, the matrices $\mathcal{M}_{21}$ and $\mathcal{M}_{22}$, for this case, are given by,

$$\mathcal{M}_{21} = \begin{array}{c} \\ 0 \\ 1 \\ \vdots \\ M-1 \\ M \end{array} \begin{array}{c} \begin{array}{ccccc} 0 & 1 & 2 & \cdots & M \end{array} \\ \begin{bmatrix} 0 & \theta & 0 & \cdots & 0 \\ 0 & 0 & \theta & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \theta \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix} \end{array}, \quad \mathcal{M}_{22} = \begin{array}{c} \\ 0 \\ 1 \\ \vdots \\ M-1 \\ M \end{array} \begin{array}{c} \begin{array}{ccccc} 0 & 1 & \cdots & M-1 & M \end{array} \\ \begin{bmatrix} 1-\theta & 0 & \cdots & 0 & 0 \\ 0 & 1-\theta & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & 0 \\ 0 & 0 & \cdots & 1-\theta & 0 \\ 0 & 0 & \cdots & 0 & 1-\theta \end{bmatrix} \end{array}.$$

We note that, if queue information is withheld from the app users, then there is always a finite probability with which the queue length becomes arbitrarily large. Therefore, the value of M, in this case, represents some arbitrarily large number which results from truncation, in order to maintain computational tractability.

Services. Finally, the intermediate transition from (x_3, u_3) to (x_0, u_0) occurs due to services. We denote the probability of generating k service slots by, $\mathbb{P}(S_t = k) = s_k$. We note that the second component of the state space, u_3, remains unchanged in this intermediate transition. Thus, $(x_3, 0)$ transitions to $(x_0, 0)$, and $(x_3, 1)$ transitions to $(x_0, 1)$. For brevity, we adopt the notation $\sum_{i=k}^{\infty} s_i = \bar{s}_k$. The probability matrix for this

intermediate transition is given by,

$$\mathcal{P}_3 = \begin{array}{c} \\ (x_3,0) \\ (x_3,1) \end{array} \begin{array}{c} \begin{array}{cc} (x_0,0) & (x_0,1) \end{array} \\ \begin{bmatrix} \mathcal{M}_3 & \mathbf{0} \\ \mathbf{0} & \mathcal{M}_3 \end{bmatrix} \end{array},$$

where $\mathbf{0}$ in the matrix $\mathcal{P}_3$ denotes a matrix containing zeros, having same dimension as that of $\mathcal{M}_3$ where $\mathcal{M}_3$ is given by,

$$\mathcal{M}_3 = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ \vdots \\ M \end{array} \begin{array}{c} \begin{array}{ccccc} 0 & 1 & 2 & \cdots & M \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ \overline{s}_1 & s_0 & 0 & \cdots & 0 \\ \overline{s}_2 & s_1 & s_0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \overline{s}_M & s_{M-1} & s_{M-2} & \cdots & s_0 \end{bmatrix} \end{array}.$$

Steady State Probability. For calculating the steady state probabilities corresponding to the system state (x_i, u_i), the transition probability matrix is given by the following product of the matrices corresponding to the intermediate transitions,

$$\overline{\mathcal{P}}_i = \mathcal{P}_{m(i)}\mathcal{P}_{m(i+1)}\mathcal{P}_{m(i+2)}\mathcal{P}_{m(i+3)},$$

where $i = 0, 1, 2$ or 3, and $m(k)$ denotes the modulo operation, which in this case, equals the remainder left when k is divided by 4. For example, if we define the system state as (x_2, u_2), then $\overline{\mathcal{P}}_2 = \mathcal{P}_2\mathcal{P}_3\mathcal{P}_0\mathcal{P}_1$. We denote the steady state probability vector corresponding to the state (x_i, u_i) by π_i, which is given by the solution of the following

equation,

$$\pi_i = \pi_i \overline{\mathcal{P}}_i. \tag{B.5}$$

The dimensions of π_i and $\overline{\mathcal{P}}_i$ are $(1, 2(M+1))$ and $(2(M+1), 2(M+1))$ respectively. The j^{th} component of π_i represents

$$\pi_i(j) = \begin{cases} \mathbb{P}(x_i = j, u_i = 0) & 0 \le j \le M, \\ \mathbb{P}(x_i = j - M - 1, u_i = 1) & M + 1 \le j \le 2M + 1. \end{cases}$$

Throughput. Using the steady state probabilities π_i, we compute the throughput for app users and non-app users. Since we have one app user arriving at the market each time period, arrival rate for app users is given by $\Lambda_T = 1$. For the model with information, throughput from online orders is given by,

$$\begin{aligned} \lambda_o &= \Lambda_T \cdot \mathbb{P}(\tau_l \le x_2 \le \tau_u) \\ &= \sum_{k=\tau_l}^{\tau_u} (\pi_2(k) + \pi_2(k + M + 1)) \\ &= \sum_{k=\tau_l}^{\tau_u} \pi_2(k) \quad (\text{since } u_2 = 0). \end{aligned}$$

For the model without information, throughput from online orders is given by,

$$\begin{aligned} \lambda_o &= \Lambda_T \cdot \mathbb{P}(\text{app user orders online}) \\ &= \theta. \end{aligned}$$

The throughput from offline (in-store) orders by app-users is given by,

$$\begin{aligned}\lambda_s &= \Lambda_T \cdot \mathbb{P}(\{0 \le x_1 < \tau_s\} \cap \{u_1 = 1\}) \\ &= \sum_{k=0}^{\tau_s - 1} \pi_1(k + M + 1).\end{aligned}$$

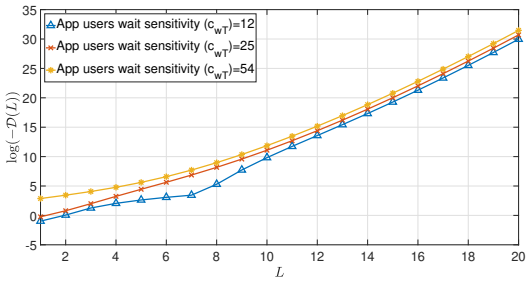
Thus, the overall throughput combining both online and offline orders by app users, is given by, $\lambda_T = \lambda_o + \lambda_s$.

Effective number of non-app users who join the system depends on the number of arrivals, A_t, the length of the queue that they observe upon arrival, x_0, and their joining threshold, τ_n. The effective number of non-app users who join the system is given by, $\min(A_t, (\tau_n - x_0)^+)$, where z^+ denotes $\max(z, 0)$. Thus, the throughput from the orders by non-app users is given by,

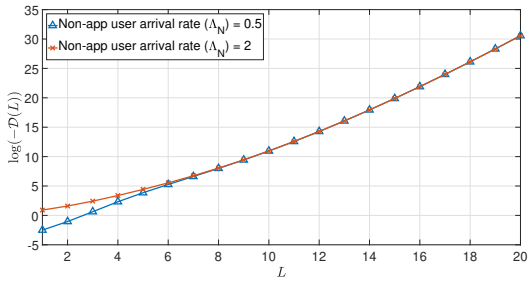
$$\begin{aligned}\lambda_N &= \mathbb{E}_{x_0, A_t} \min(A_t, (\tau_n - x_0)^+) \\ &= \sum_{k=0}^{\tau_n - 1} \sum_{j=0}^{\infty} \mathbb{P}(A_t = j) \cdot (\pi_0(k) + \pi_0(k + M + 1)) \cdot \min(j, (\tau_n - k)^+).\end{aligned}$$

Consumer Surplus. Using the steady state probabilities, we compute the average per period consumer surplus for the app users and non-app users. For the model with information, average consumer surplus for app users is computed as follows,

$$\begin{aligned}C_T &= \mathbb{E}_{x_2} \max(U_s(x_2), U_o(x_2)) \\ &= \sum_{l=0}^{M} (\pi_2(l) + \pi_2(l + M + 1)) \cdot \max(U_s(l), U_o(l)) \\ &= \sum_{l=0}^{M} \pi_2(l) \cdot \max(U_s(l), U_o(l)) \quad (\text{since } u_2 = 0),\end{aligned}$$



(a) $c_{wN} = 30,\ \Lambda_N = 0.9$



(b) $c_{wN} = 20,\ c_{wT} = 22$

Figure B.1. Plot (a) and (b) illustrates that Assumption 2 holds. For scaling purposes we plot $\log(-\mathcal{D}(L))$ on the y-axis, which is a monotone transform of $-\mathcal{D}(L)$. If $\log(-\mathcal{D}(L))$ is increasing in L that implies $\mathcal{D}(L)$ is decreasing in L. These graphs are plotted for base parameter values used in §2.5.

where $U_o(l)$ and $U_s(l)$ are defined in (2.1) and (2.2). For the model without information, average consumer surplus for app users is computed as follows,

$$C_T = \max(\mathbb{E}_{x_2} U_s(x_2), \mathbb{E}_{x_2} U_o(x_2)).$$

For non-app users, we have established that, the effective number of customers joining the system is given by $n_t = \min(A_t, (\tau_n - x_0)^+)$. First we compute the combined consumer surplus for all joining non-app users conditional on n_t, which is given as follows,

$$c(A_t, x_0) = \sum_{i=1}^{n_t} \left(v - (x_0 + i) \cdot \frac{c_{wN}}{\mu} \right).$$

Now, the average per period consumer surplus for non-app users is given by,

$$\begin{aligned} C_N &= \mathbb{E}_{x_0, A_t} c(A_t, x_0) \\ &= \sum_{k=0}^{M} \sum_{j=0}^{\infty} \mathbb{P}(A_t = j) \cdot (\pi_0(k) + \pi_0(k + M + 1)) \cdot c(j, k). \end{aligned}$$

Now that we have presented the framework which we will use for computations in §2.5, we illustrate in Figure B.1 that Assumption 2 holds for the base parameter values that we consider in §2.5.

B.3. Throughput for Low System Capacity

Here, we will show that there exist a capacity, μ, such that app users arriving at the market order online after observing an empty system, even if no customer joins the queue in store. In particular, we assume that $v - \frac{c_{wT}}{\mu} < 0$. Since, in this setting, non-app users are more wait-sensitive compared to app users, we have $v - \frac{c_{wN}}{\mu} < 0$ as well. Thus, neither non-app users nor app users join in store even if the queue is empty. In addition, we assume that app users are not quality sensitive, i.e. $c_q = 0$. Now, we consider app users' utility for choosing the offline option upon observing an empty system. First, from (2.3), we have,

$$\hat{U}_s(0) = v - \Big((0 - S_t)^+ + n_{t+1}(0) + 1\Big) \cdot \frac{c_{wT}}{\mu}.$$

Since, non-app users do not join , i.e., $v - \frac{c_{wN}}{\mu} < 0$, we have $n_{t+1}(0) = 0$. Thus, we have,

$$\hat{U}_s(0) = v - \Big((0 - S_t)^+ + 1\Big) \cdot \frac{c_{wT}}{\mu} = v - \frac{c_{wT}}{\mu}.$$

Since, we have assumed that $v - \frac{c_{wT}}{\mu} < 0$, we have $\hat{U}_s(0) < 0$, and hence $U_s(0) = 0$ from (2.2). Thus app users do not choose the offline option. Now, we consider the utility of ordering online. From (2.1), we have,

$$U_o(0) = v - \mathbb{E}_{S_t}[(0 + 1 - S_t)^+] \cdot \frac{c_{wT}}{\mu} = v - \mathbb{E}_{S_t}[(1 - S_t)^+] \cdot \frac{c_{wT}}{\mu}.$$

Since, $\mathbb{E}_{S_t}[(1-S_t)^+] < 1$ for any non-zero service process, there exist a capacity, μ, such that $\mathbb{E}_{S_t}[(1-S_t)^+] \cdot \frac{c_{wT}}{v} < \mu < \frac{c_{wT}}{v}$, which results in $U_o(0) > 0$. Thus, for low enough capacity, even though in-store throughput could be zero, app users still might order online resulting in positive omnichannel throughput. ■

B.4. Proofs of Results in §2.3

Proof of Theorem 4

A strategy, $\theta \in [0,1]$, denotes the probability with which an app user arriving at the market orders online. Before we establish the existence of a symmetric equilibrium, we look at the best response, which is denoted by the correspondence, $\chi(\theta) : [0,1] \to [0,1]$, and is given by,

$$\chi(\theta) \begin{cases} = 1, & \text{if } \overline{U}_o(\theta) > \overline{U}_s(\theta) \\ \in [0,1], & \text{if } \overline{U}_o(\theta) = \overline{U}_s(\theta) \\ = 0, & \text{if } \overline{U}_o(\theta) < \overline{U}_s(\theta) \end{cases} \tag{B.6}$$

where $\overline{U}_o(\theta)$ and $\overline{U}_s(\theta)$ are given by (2.7). In order to prove the existence of a symmetric equilibrium strategy, it is sufficient to prove the existence of a fixed point for the correspondence $\chi(\theta) : [0,1] \to [0,1]$. An equilibrium strategy, θ, would be such that $\theta = \chi(\theta)$. The set of all strategies, $[0,1]$, is a closed and compact set. Clearly, the correspondence $\chi(\theta)$, as defined in (B.6), is convex-valued. To see this, note that for any two elements $r_1, r_2 \in \chi(\theta)$ and $r_1 \neq r_2$, by definition of $\chi(\theta)$, θ is such that $\overline{U}_o(\theta) = \overline{U}_s(\theta)$. Therefore, for any $\alpha \in [0,1]$, $r = \alpha r_1 + (1-\alpha) r_2 \in [0,1]$, and hence $r \in \chi(\theta)$. If $r_1 = r_2$, then clearly, $r \in \chi(\theta)$. Now, the steady state probabilities π_2, are given by the solution of the equation

(B.5), where $\overline{\mathcal{P}}_2 = \mathcal{P}_2\mathcal{P}_3\mathcal{P}_0\mathcal{P}_1$ equals the matrix $\mathcal{M}_{21}\mathcal{M}_3\mathcal{M}_0 + \mathcal{M}_{22}\mathcal{M}_3\mathcal{M}_0\mathcal{M}_1$, where, only matrices $\mathcal{M}_{21}$ and $\mathcal{M}_{22}$ are functions of θ. Thus, (B.5) represents a set of equations that are linear in θ, and hence continuous in θ. As a result, the expected utility functions given by, (2.7), are also continuous functions of θ. The best response correspondence $\chi(\theta)$, as defined in (B.6), is continuous in the function $\overline{U}_o(\theta) - \overline{U}_s(\theta)$. Since, we have established that both $\overline{U}_o(\theta)$ and $\overline{U}_s(\theta)$ are continuous in θ, this implies that $\chi(\theta)$ is also continuous in θ, by composition of $\chi(\theta)$ with $\overline{U}_o(\theta)$ and $\overline{U}_s(\theta)$. Thus the best response correspondence, $\chi(\theta)$, has a closed graph, i.e., the set $\{(\theta, \phi) \in [0,1]^2 : \phi \in \chi(\theta)\}$ is closed as a subset of $[0,1]^2$. Hence, by Kakutani's fixed point theorem, the best response correspondence, $\chi(\theta)$, has a fixed point θ such that $\theta = \chi(\theta)$. This is a symmetric Nash equilibrium, as it is the best response for an app user arriving at the market, when all other app users who arrive at the market also use the same strategy. ■

Proof of Proposition 4

Utility of ordering online is,

$$U_o(L) = v - \mathbb{E}_{S_t}[\mathbb{1}(S_t \geq L+1)] \cdot c_q - \mathbb{E}_{S_t}[(L+1-S_t)^+] \cdot \frac{c_{wT}}{\mu}.$$

Utility of choosing the offline option is,

$$U_s(L) = \mathbb{E}_{S_t, A_{t+1}}[\max(0, \hat{U}_s(L))],$$

where,

$$\hat{U}_s(L) = v - \Big((L - S_t)^+ + n_{t+1}(L) + 1\Big) \cdot \frac{c_{wT}}{\mu}.$$

Since the function $f(x) = \max(0, x)$ is convex in x, using Jensen's inequality we have,

$$\mathbb{E}_{S_t, A_{t+1}}[\max(0, \hat{U}_s(L))] \geq \max(0, \mathbb{E}_{S_t, A_{t+1}}[\hat{U}(L)]).$$

Since, $\max(0, \mathbb{E}_{S_t, A_{t+1}}[\hat{U}(L)]) \geq \mathbb{E}_{S_t, A_{t+1}}[\hat{U}_s(L)]$, we have $U_s(L) \geq \mathbb{E}_{S_t, A_{t+1}}[\hat{U}_s(L)]$. Hence, we will express $U_s(L)$ as $\mathbb{E}_{S_t, A_{t+1}}[\hat{U}_s(L)] + g(L)$, where $g(L)$ is a positive function representing the gap, and is defined over the set of all non-negative integers. The difference, $D(L) := U_o(L) - U_s(L)$, can be written as,

$$\begin{aligned}
D(L) &= U_o(L) - \mathbb{E}_{S_t, A_{t+1}}[\hat{U}_s(L)] - g(L) \\
&= \mathbb{E}_{S_t}[(L - S_t)^+ + \mathbb{E}_{A_{t+1}} n_{t+1}(L) + 1] \cdot \frac{c_{wT}}{\mu} \\
&\quad - \mathbb{E}_{S_t}[\mathbb{1}(S_t \geq L + 1)] \cdot c_q - \mathbb{E}_{S_t}[(L + 1 - S_t)^+] \cdot \frac{c_{wT}}{\mu} - g(L) \\
&= \mathbb{E}_{S_t, A_{t+1}}[n_{t+1}(L)] \cdot \frac{c_{wT}}{\mu} + \mathbb{E}_{S_t}[(L - S_t)^+ + 1 - (L + 1 - S_t)^+] \cdot \frac{c_{wT}}{\mu} \\
&\quad - \mathbb{P}(S_t \geq L + 1) \cdot c_q - g(L),
\end{aligned}$$

which is equivalent to,

$$D(L) = \mathbb{E}_{S_t, A_{t+1}}[n_{t+1}(L)] \cdot \frac{c_{wT}}{\mu} + \mathbb{P}(S_t \geq L + 1) \cdot \left(\frac{c_{wT}}{\mu} - c_q \right) - g(L). \tag{B.7}$$

We are interested in finding out the number of times, the function $D(L)$ crosses zero. In order to investigate this, it would be helpful to look at the first differences of this function, $D(L) - D(L-1)$. The first difference function would tell us how $D(L)$ changes

as L changes. Computing $D(L) - D(L-1)$, we have,

$$D(L) - D(L-1) = \mathbb{E}_{S_t, A_{t+1}}[n_{t+1}(L) - n_{t+1}(L-1)] \cdot \frac{c_{wT}}{\mu} - \mathbb{P}(S_t = L) \cdot \Big(\frac{c_{wT}}{\mu} - c_q\Big) - (g(L) - g(L-1)).$$

Therefore, the condition $D(L) < D(L-1)$ is equivalent to

$$\frac{\mathbb{E}_{S_t, A_{t+1}}[n_{t+1}(L) - n_{t+1}(L-1)] \cdot \frac{c_{wT}}{\mu} - (g(L) - g(L-1))}{\mathbb{P}(S_t = L)} < \Big(\frac{c_{wT}}{\mu} - c_q\Big). \tag{B.8}$$

As per Assumption 2, the left hand side expression in (B.8) is $\mathcal{D}(L)$, and we assume that $\mathbb{P}(S_t = L) > 0$ for all non-negative integer L. Now, given that $n_{t+1}(L) = \min(A_{t+1}, (\tau_n - (L - S_t)^+)^+)$, clearly, $n_{t+1}(L) \leq n_{t+1}(L-1)$, which implies that the term $\mathbb{E}_{S_t, A_{t+1}}[n_{t+1}(L) - n_{t+1}(L-1)] \cdot \frac{c_{wT}}{\mu}$, in (B.8) is non-positive. Moreover Lemma 8 establishes that the term $(g(L) - g(L-1))$ in (B.8) is positive. This implies that $\mathcal{D}(L)$ is negative for all L.

(i) If $\frac{c_{wT}}{\mu} - c_q \geq 0$, then (B.8) holds for all $L \geq 0$. This implies that $D(L) < D(L-1)$ for all $L \geq 1$. Now, to avoid trivialities we generally assume that $\tau_s \geq 1$, i.e. $v - \frac{c_{wT}}{\mu} \geq 0$. Lemma 10 establishes that $D(0) \geq 0$. Hence, there exists a threshold τ_u, such that $D(L) \geq 0$ if $L \leq \tau_u$ and $D(L) < 0$ otherwise. This proves the first part of the proposition.

(ii) If $\frac{c_{wT}}{\mu} - c_q < 0$ and Assumption 2 holds, then either $\mathcal{D}(L) < \big(\frac{c_{wT}}{\mu} - c_q\big)$ for all $L \geq 0$, i.e. (B.8) holds for all $L \geq 0$ which we have covered in (i), or there exists an L^* such that $\big(\frac{c_{wT}}{\mu} - c_q\big) \leq \mathcal{D}(L)$ if $L < L^*$ and $\mathcal{D}(L) < \big(\frac{c_{wT}}{\mu} - c_q\big)$ otherwise. This implies that once $D(L) < D(L-1)$ holds for some L^*, it continues to hold true for all subsequent $L > L^*$, and hence the function $D(L) = U_o(L) - U_s(L)$ is unimodal. Thus, if $D(L) \geq 0$ for some

$L \geq 0$, there exist thresholds $\tau_l \geq 0$ and $\tau_u \geq 0$ such that $U_o(L) \geq U_s(L)$ if $\tau_l \leq L \leq \tau_u$, and $U_o(L) < U_s(L)$ otherwise. ■

Lemma 8. *The function $g(L)$ defined as,*

$$g(L) := \mathbb{E}(\max(0, \hat{U}_s(L))) - \mathbb{E}(\hat{U}_s(L))$$

is increasing in L.

Proof. From (2.3), we have,

$$\hat{U}_s(L) = v - ((L - S_t)^+ + n_{t+1}(L) + 1) \cdot \frac{c_{wT}}{\mu}.$$

We can rewrite this as,

$$\hat{U}_s(L) = v - X_L \cdot \frac{c_{wT}}{\mu},$$

where $X_L = (L - S_t)^+ + n_{t+1}(L) + 1$ is a discrete positive random variable. $\mathbb{E}(\max(0, \hat{U}_s(L)))$ can be written as,

$$\begin{aligned} \mathbb{E}(\max(0, \hat{U}_s(L))) &= \mathbb{E}\Big(\max\Big(0, v - X_L \cdot \frac{c_{wT}}{\mu}\Big)\Big) \\ &= v - \frac{c_{wT}}{\mu} \cdot \mathbb{E}\Big(\min\Big(\frac{v\mu}{c_{wT}}, X_L\Big)\Big) \\ &= v - \frac{v}{\gamma} \cdot \mathbb{E}(\min(\gamma, X_L)), \end{aligned} \tag{B.9}$$

where we define $\gamma = \frac{v\mu}{c_{wT}}$.

Expanding the term $\mathbb{E}(\min(\gamma, X_L))$, we get,

$$\begin{aligned}\mathbb{E}(\min(\gamma, X_L)) &= \sum_{i=1}^{\lfloor\gamma\rfloor} i \cdot \mathbb{P}(X_L = i) \\ &= \sum_{i=0}^{\lfloor\gamma\rfloor} \mathbb{P}(\gamma > X_L > i).\end{aligned} \tag{B.10}$$

Now, we expand the term $\mathbb{E}(\hat{U}_s(L))$ in a similar manner, we get,

$$\begin{aligned}\mathbb{E}(\hat{U}_s(L)) &= v - \frac{v}{\gamma} \cdot \mathbb{E}(X_L) \\ &= v - \frac{v}{\gamma} \cdot \sum_{i=0}^{\infty} \mathbb{P}(X_L > i).\end{aligned} \tag{B.11}$$

Finally, combining (B.9), (B.10) and (B.11), we can write,

$$\begin{aligned}g(L) &= \frac{v}{\gamma} \cdot \Big(\sum_{i=0}^{\infty} \mathbb{P}(X_L > i) - \sum_{i=0}^{\lfloor\gamma\rfloor} \mathbb{P}(\gamma > X_L > i) \Big) \\ &= \frac{v}{\gamma} \cdot \Big(\sum_{i=\lfloor\gamma\rfloor+1}^{\infty} \mathbb{P}(X_L > i) + \sum_{i=0}^{\lfloor\gamma\rfloor} \mathbb{P}(X_L > \gamma) \Big).\end{aligned} \tag{B.12}$$

Using Lemma 9, we can infer that $\mathbb{P}(X_L > i)$ is increasing in L for any i. This establishes our result that $g(L)$ is increasing in L. ■

Lemma 9. *X_L is stochastically increasing in L.*

Proof. The discrete positive random variable, X_L, is defined as $X_L = (L - S_t)^+ + n_{t+1}(L) + 1$, where $n_{t+1}(L) = \min(A_{t+1}, (\tau_n - (L - S_t)^+)^+)$.

First, considering the case $S_t \geq L+1$, we have,

$$X_L = 0 + \min(A_{t+1}, \tau_n) + 1$$

and

$$X_{L+1} = 0 + \min(A_{t+1}, \tau_n) + 1.$$

This implies, $X_{L+1} = X_L$. Now, considering $S_t \leq L$, we get,

$$X_{L+1} - X_L = (L+1-S_t) - (L-S_t) + \min(A_{t+1}, (\tau_n + S_t - L - 1)) - \min(A_{t+1}, (\tau_n + S_t - L)).$$

The term $\min(A_{t+1}, (\tau_n + S_t - L - 1)) - \min(A_{t+1}, (\tau_n + S_t - L))$, is either 0 or -1, which implies $X_{L+1} - X_L$ is either 1 or 0. This implies that $X_{L+1} \geq X_L$ and thus, X_L is stochastically increasing in L. ■

Lemma 10. *$D(0) \geq 0$ if $v - \frac{c_{wT}}{\mu} \geq 0$ and $\frac{c_{wT}}{\mu} - c_q \geq 0$.*

Proof. $D(0) = U_o(0) - U_s(0)$ is equivalent to,

$$D(0) = v - \mathbb{E}_{S_t}[\mathbb{1}(S_t \geq 0+1)] \cdot c_q - \mathbb{E}_{S_t}[(0+1-S_t)^+] \cdot \frac{c_{wT}}{\mu} - \mathbb{E}_{S_t, A_{t+1}}[\max(0, \hat{U}_s(0))] \tag{B.13}$$

where,

$$\hat{U}_s(0) = v - \Big((0 - S_t)^+ + n_{t+1}(0) + 1\Big) \cdot \frac{c_{wT}}{\mu} = v - \Big(n_{t+1}(0) + 1\Big) \cdot \frac{c_{wT}}{\mu}. \tag{B.14}$$

Now, substituting $\hat{U}_s(0)$ from (B.14) in $\mathbb{E}_{S_t, A_{t+1}}[\max(0, \hat{U}_s(0))]$, we can rewrite (B.13) as,

(B.15)

$$D(0) = v - \mathbb{P}(S_t \geq 1) \cdot c_q - (1 - \mathbb{P}(S_t \geq 1)) \cdot \frac{c_{wT}}{\mu} - \mathbb{E}_{S_t, A_{t+1}}\Big[\max\Big(0, v - (n_{t+1}(0)+1) \cdot \frac{c_{wT}}{\mu}\Big)\Big].$$

Since $\max\left(0, v-(n_{t+1}(0)+1)\cdot\frac{c_{wT}}{\mu}\right) \leq \max\left(0, v-\frac{c_{wT}}{\mu}\right)$, we have the following inequality which follows from (B.15),

$$D(0) \geq \left(v - \frac{c_{wT}}{\mu}\right) + \mathbb{P}(S_t \geq 1)\cdot\left(\frac{c_{wT}}{\mu} - c_q\right) - \max\left(0, v - \frac{c_{wT}}{\mu}\right). \tag{B.16}$$

If $v - \frac{c_{wT}}{\mu} \geq 0$ and $\frac{c_{wT}}{\mu} - c_q \geq 0$, from (B.16) we have $D(0) \geq 0$. ■

B.5. Proofs of Results in §2.4

Here, we present the proofs for the propositions presented in §2.4. We denote $\mathbb{P}(A_t = 0)$ by a and $\mathbb{P}(S_t = 1)$ by s. In both the Patient and the Impatient Scenarios, $a = \frac{1}{2}$ and $s = \frac{1}{2}$. For deriving the analytical expressions for the steady state probabilities, throughput and consumer surplus in the Patient and the Impatient scenarios, we use the framework presented in Appendix B.2. For both the Patient Scenario and the Impatient Scenario, we have

$$\mathcal{M}_0 = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \end{array}\!\!\begin{array}{c} \begin{array}{cccc} 0 & 1 & 2 & 3 \end{array} \\ \begin{bmatrix} a & 1-a & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \end{array}, \qquad \mathcal{M}_3 = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \end{array}\!\!\begin{array}{c} \begin{array}{cccc} 0 & 1 & 2 & 3 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1-s & s & 0 & 0 \\ 0 & 1-s & s & 0 \end{bmatrix} \end{array}. \tag{B.17}$$

Moreover, we have

$$\text{(B.18)} \qquad \mathcal{M}_1 = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{cccc} 0 & 1 & 2 & 3 \end{array} \\ \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \end{array}, \qquad \mathcal{M}_1 = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{cccc} 0 & 1 & 2 & 3 \end{array} \\ \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \end{array}$$

for the Patient and the Impatient Scenarios respectively.

For the model without information, we have

$$\text{(B.19)} \qquad \mathcal{M}_{21} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{cccc} 0 & 1 & 2 & 3 \end{array} \\ \begin{bmatrix} 0 & t & 0 & 0 \\ 0 & 0 & t & 0 \\ 0 & 0 & 0 & t \\ 0 & 0 & 0 & 0 \end{bmatrix} \end{array}, \qquad \mathcal{M}_{22} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{cccc} 0 & 1 & 2 & 3 \end{array} \\ \begin{bmatrix} 1-t & 0 & 0 & 0 \\ 0 & 1-t & 0 & 0 \\ 0 & 0 & 1-t & 0 \\ 0 & 0 & 0 & 1-t \end{bmatrix} \end{array}$$

in both the Patient and the Impatient Scenarios, where we use t to denote the probability that an app user orders online. For single channel systems, $\mathcal{M}_{21}$ and $\mathcal{M}_{22}$ can be obtained by putting $t = 0$ in (B.19), since all app users choose the offline option. For the model with information, we will specify the matrices $\mathcal{M}_{21}$ and $\mathcal{M}_{22}$ in the proof of the corresponding Lemmas. For the remainder of this section, we adopt the notation $\hat{\pi}_i(k,l) = \mathbb{P}(x_i = k, u_i = l)$, and $\overline{\pi}_i(k) = \hat{\pi}_i(k,0) + \hat{\pi}_i(k,1)$.

Proof of Proposition 5

The result follows from Lemma 17. ■

Proof of Proposition 6

The result follows from Lemma 11. ■

Proof of Proposition 7

The result follows from Lemma 18, Lemma 20 and Lemma 22 combined. ■

Proof of Proposition 8

The result follows from Lemma 12, Lemma 14 and Lemma 16 combined. ■

Patient Scenario

System: Omnichannel without information

Lemma 11. *(i) The probability, θ, that an app user arriving at the market orders online, in the omnichannel system without information for the Patient Scenario, is given by,*

$$\theta = \begin{cases} 1 & \text{if } 0 < \dfrac{c_q}{c_{wT}} \leq \dfrac{7}{6}, \\ \dfrac{6c_{wT} - 3c_q}{3c_q - c_{wT}} & \text{if } \dfrac{7}{6} < \dfrac{c_q}{c_{wT}} < 2, \\ 0 & \text{if } 2 < \dfrac{c_q}{c_{wT}} < \infty. \end{cases}$$

(ii) The average per period consumer surplus for the app users in the omnichannel system

without information for the Patient Scenario, is given by,

$$C_T = \begin{cases} v - \frac{2}{3}c_q - \frac{2}{9}c_{wT} & \text{if } 0 < \frac{c_q}{c_{wT}} \leq \frac{7}{6}, \\ \frac{36vc_q - 42c_qc_{wT} - 27vc_{wT} + 34c_{wT}^2}{36c_q - 27c_{wT}} & \text{if } \frac{7}{6} < \frac{c_q}{c_{wT}} < 2, \\ v - \frac{10}{9}c_{wT} & \text{if } 2 \leq \frac{c_q}{c_{wT}} < \infty. \end{cases}$$

where $\frac{4}{3} \leq \frac{v}{c_{wT}} < 2$.

(iii) The average per period consumer surplus for the non-app users in the omnichannel system without information for the Patient Scenario, is given by,

$$C_N = \begin{cases} \frac{1}{3} \cdot \left(v - \frac{2}{3}c_{wN}\right) & \text{if } 0 < \frac{c_q}{c_{wT}} \leq \frac{7}{6}, \\ \frac{1}{2} \cdot \frac{36c_q^2 - 72c_qc_{wT} + 40c_{wT}^2}{36c_q^2 - 51c_qc_{wT} + 18c_{wT}^2} \cdot \left(v - \frac{2}{3}c_{wN}\right) & \text{if } \frac{7}{6} < \frac{c_q}{c_{wT}} < 2, \\ \frac{1}{3} \cdot \left(v - \frac{2}{3}c_{wN}\right) & \text{if } 2 \leq \frac{c_q}{c_{wT}} < \infty. \end{cases}$$

where $\frac{2}{3} \leq \frac{v}{c_{wN}} < \frac{4}{3}$.

Proof. The steady state probabilities when an app user arrives at the market, are given by,

$$\overline{\pi}_2(0) = \frac{(t^2+t)}{(t-3)(t-4)}, \quad \overline{\pi}_2(1) = \frac{2(2+t-t^2)}{(t-3)(t-4)}, \quad \overline{\pi}_2(2) = \frac{2(t-1)}{(t-3)}, \tag{B.20}$$

where t denotes the probability that an app user arriving at the market orders online. Next, we compute the expected utilities of ordering online and choosing the offline option, for an app user arriving at the market. We recall that the utility of ordering online from

(2.1) is given by,

$$U_o(L) = v - \mathbb{E}_{S_t}[\mathbb{1}(S_t \geq L+1)] \cdot c_q - \mathbb{E}_{S_t}[(L+1-S_t)^+] \cdot \frac{c_{wT}}{\mu}.$$

Thus, we have,

$$\begin{aligned} U_o(0) &= v - \mathbb{E}_{S_t}[\mathbb{1}(S_t \geq 1)] \cdot c_q - \mathbb{E}_{S_t}[(1-S_t)^+] \cdot \frac{c_{wT}}{\mu} \\ &= v - c_q. \end{aligned} \tag{B.21}$$

$$\begin{aligned} U_o(1) &= v - \mathbb{E}_{S_t}[\mathbb{1}(S_t \geq 2)] \cdot c_q - \mathbb{E}_{S_t}[(2-S_t)^+] \cdot \frac{c_{wT}}{\mu} \\ &= v - s \cdot c_q - (1-s) \cdot \frac{c_{wT}}{\mu} \\ &= v - \frac{c_q}{2} - \frac{c_{wT}}{3}. \qquad \left[\text{Note: } \mu = 1 \cdot s + 2 \cdot (1-s) = \frac{3}{2}\right] \end{aligned} \tag{B.22}$$

$$\begin{aligned} U_o(2) &= v - \mathbb{E}_{S_t}[\mathbb{1}(S_t \geq 3)] \cdot c_q - \mathbb{E}_{S_t}[(3-S_t)^+] \cdot \frac{c_{wT}}{\mu} \\ &= v - [2 \cdot (1-s) + 1 \cdot s] \cdot \frac{c_{wT}}{\mu} \\ &= v - c_{wT}. \end{aligned} \tag{B.23}$$

Now, in the Patient Scenario we have,

$$2 \leq \frac{v\mu}{c_{wT}} < 3$$

which, since $\mu = 1 \cdot s + 2 \cdot (1-s) = \frac{3}{2}$, implies

$$\frac{4}{3} \leq \frac{v}{c_{wT}} < 2. \tag{B.24}$$

The utility of choosing the offline option from (2.2) is given by,

$$U_s(L) = \mathbb{E}_{S_t, A_{t+1}}[\max(0, \hat{U}_s(L))]$$

where $\hat{U}_s(L)$ is given by,

$$\hat{U}_s(L) = v - ((L - S_t)^+ + n_{t+1}(L) + 1) \cdot \frac{c_{wT}}{\mu}.$$

Thus, we have,

$$U_s(0) = \mathbb{E}_{S_t, A_{t+1}}[\max(0, v - (n_{t+1}(0) + 1) \cdot \frac{c_{wT}}{\mu})].$$

We know that $n_{t+1}(L) = \min(A_{t+1}, (\tau_n - (L - S_t)^+)^+)$. Thus, $n_{t+1}(0) = \min(A_{t+1}, \tau_n)$. Since, we have assumed that $\tau_n = 1$, we have, $n_{t+1}(0) = A_{t+1}$. Using this, we can rewrite (B.25) as,

$$U_s(0) = \mathbb{E}_{A_{t+1}}[\max(0, v - (A_{t+1} + 1) \cdot \frac{c_{wT}}{\mu})].$$

Since, A_{t+1} is either 0 with probability $a = \frac{1}{2}$, or 1 with probability $1 - a = \frac{1}{2}$, and (B.24) holds, we can rewrite $U_s(0)$ as,

$$U_s(0) = (v - 2\frac{c_{wT}}{\mu}) \cdot a + (v - \frac{c_{wT}}{\mu}) \cdot (1 - a) = v - c_{wT}. \tag{B.25}$$

Since, $n_{t+1}(1) = \min(A_{t+1}, (\tau_n - (1 - S_t)^+)^+) = \min(A_{t+1}, \tau_n) = A_{t+1}$, we have,

$$\begin{aligned} U_s(1) &= \mathbb{E}_{A_{t+1}}[\max(0, v - (A_{t+1} + 1) \cdot \frac{c_{wT}}{\mu})] \\ &= v - c_{wT}. \end{aligned} \tag{B.26}$$

Finally,

$$
\begin{aligned}
n_{t+1}(2) &= \min(A_{t+1}, (\tau_n - (2 - S_t)^+)^+) \\
&= \min(A_{t+1}, (1 - (2 - S_t)^+)^+) \\
&= \begin{cases} 0 & \text{if } S_t = 1, \\ A_{t+1} & \text{if } S_t = 2. \end{cases}
\end{aligned}
$$

Thus,

$$
\begin{aligned}
U_s(2) &= \mathbb{E}_{S_t, A_{t+1}}[\max(0, v - ((2 - S_t)^+ + n_{t+1}(2) + 1) \cdot \frac{c_{wT}}{\mu})] \\
&= \begin{cases} \mathbb{E}_{A_{t+1}}[\max(0, v - 2\frac{c_{wT}}{\mu})] \text{ if } S_t = 1, \\ \mathbb{E}_{A_{t+1}}[\max(0, v - (A_{t+1} + 1) \cdot \frac{c_{wT}}{\mu})] \text{ if } S_t = 2, \end{cases} \\
&= \begin{cases} v - \frac{4}{3} c_{wT}, \text{ if } S_t = 1, \\ v - c_{wT} \text{ if } S_t = 2, \end{cases} \\
&= (1 - s) \cdot (v - \frac{4}{3} c_{wT}) + s \cdot (v - c_{wT}) \\
&= v - \frac{7}{6} c_{wT}.
\end{aligned}
\tag{B.27}
$$

Collecting all the utility expression from (B.21), (B.22), (B.23), (B.25), (B.26) and (B.27) we have the utility expressions in Table B.1.

We next compute the expected utilities of ordering online and choosing the offline option by app users arriving at the market, using steady state probabilities given by (B.20).

L	$U_o(L)$	$U_s(L)$
0	$v - c_q$	$v - c_{wT}$
1	$v - \frac{c_q}{2} - \frac{c_{wT}}{3}$	$v - c_{wT}$
2	$v - c_{wT}$	$v - \frac{7}{6}c_{wT}$

Table B.1. Utilities of ordering online and choosing the offline option by an app user when there are L order in the system

First, we compute the expected utility of ordering online,

$$\overline{U}_o(t) = \frac{6c_q - 36v + 28c_{wT} + 6c_q t + 21vt - 28c_{wT}t - 3vt^2 + 4c_{wT}t^2}{3(7t - t^2 - 12)}. \tag{B.28}$$

Expected utility of choosing the offline option is given by,

$$\overline{U}_s(t) = \frac{9v - 10c_{wT} - 3vt + 4c_{wT}t}{3(3 - t)}. \tag{B.29}$$

Now, the app users randomize their choice of channel only if they are indifferent between the online and the offline option. Thus, if $0 < t < 1$, then we have, $\overline{U}_o(t) = \overline{U}_s(t)$. Equating (B.28) and (B.29) we get,

$$t = \frac{3c_q - 6c_{wT}}{c_{wT} - 3c_q}. \tag{B.30}$$

Applying the condition $0 < t < 1$, we get,

$$2 > \frac{c_q}{c_{wT}} > \frac{7}{6}. \tag{B.31}$$

Finally, we characterize the average per period consumer surplus for app users as follows:

$$C_T = \begin{cases} v - \frac{2}{3}c_q - \frac{2}{9}c_{wT} & \text{if } 0 < \frac{c_q}{c_{wT}} \leq \frac{7}{6}, \\ \frac{36vc_q - 42c_qc_{wT} - 27vc_{wT} + 34c_{wT}^2}{36c_q - 27c_{wT}} & \text{if } \frac{7}{6} < \frac{c_q}{c_{wT}} < 2, \\ v - \frac{10}{9}c_{wT} & \text{if } 2 \leq \frac{c_q}{c_{wT}} < \infty, \end{cases} \tag{B.32}$$

where the parameters need to satisfy the condition in (B.24).

Now, we compute the average consumer surplus for non-app users in this system. We recall that in this system non-app users join only if they observe an empty queue. The steady state probability is given by,

$$\overline{\pi}_0(0) = \frac{4(t^2 - 2t + 2)}{t^2 - 7t + 12}$$

where t denotes the probability of an app user ordering online. Thus, the average per period consumer surplus for non-app users is given by,

$$C_N = \overline{\pi}_0(0) \cdot \mathbb{P}(A_t = 1) \cdot (v - \frac{c_{wN}}{\mu}).$$

Substituting the expression for t, as given in (B.30), in the expression for $\overline{\pi}_0(0)$, and using the condition (B.31) along with the fact that $\mu = \frac{3}{2}$, we can rewrite C_N as,

$$C_N = \begin{cases} \frac{1}{3} \cdot \left(v - \frac{2}{3}c_{wN}\right) & \text{if } 0 < \frac{c_q}{c_{wT}} \leq \frac{7}{6}, \\ \frac{1}{2} \cdot \frac{36c_q^2 - 72c_qc_{wT} + 40c_{wT}^2}{36c_q^2 - 51c_qc_{wT} + 18c_{wT}^2} \cdot \left(v - \frac{2}{3}c_{wN}\right) & \text{if } \frac{7}{6} < \frac{c_q}{c_{wT}} < 2, \\ \frac{1}{3} \cdot \left(v - \frac{2}{3}c_{wN}\right) & \text{if } 2 \leq \frac{c_q}{c_{wT}} < \infty, \end{cases} \tag{B.33}$$

where, the non-app users' wait-sensitivity is such that,

$$\frac{2}{3} \leq \frac{v}{c_{wN}} < \frac{4}{3}. \tag{B.34}$$

■

Lemma 12. *The combined average per period throughput in the omnichannel system without information for the Patient Scenario is given by,*

$$\lambda_T + \lambda_N = \begin{cases} \frac{4}{3} & \text{if } 0 < \frac{c_q}{c_{wT}} \leq \frac{7}{6}, \\ 1 + \frac{18c_q^2 - 36c_q c_{wT} + 20c_{wT}^2}{36c_q^2 - 51c_q c_{wT} + 18c_{wT}^2} & \text{if } \frac{7}{6} < \frac{c_q}{c_{wT}} < 2, \\ \frac{4}{3} & \text{if } 2 \leq \frac{c_q}{c_{wT}} < \infty. \end{cases}$$

Proof. Steady state probability that a non-app user arriving at the store observes an empty system is,

$$\overline{\pi}_0(0) = \hat{\pi}(0,0) + \hat{\pi}(0,1) = \frac{2(t^2+t)}{(t-3)(t-4)} + \frac{2(t-1)}{(t-3)} = \frac{4(t^2-2t+2)}{t^2-7t+12},$$

where t denotes the probability that an app user arriving at the market orders online. Average per period throughput for non-app users is,

$$\begin{aligned} \lambda_N &= \mathbb{P}(A_t = 1) \cdot \overline{\pi}_0(0) \\ &= \frac{1}{2} \cdot \Big(\frac{4(t^2-2t+2)}{t^2-7t+12}\Big) = \frac{2(t^2-2t+2)}{t^2-7t+12}. \end{aligned}$$

The steady state probability of an app user arriving at the store and observing an empty system is $\hat{\pi}_1(0,1) = \frac{(t-1)}{(t-3)}$ where t denotes the probability of an app user ordering online.

The probability of an app user arriving at the store and observing a system of size 1 is $\hat{\pi}_1(1,1) = \frac{(1-t)(t-2)}{(t-3)}$. Average per period throughput for app users is,

$$\begin{aligned}\lambda_T = \lambda_o + \lambda_s &= \Lambda_T \cdot (t + \hat{\pi}_1(0,1) + \hat{\pi}_1(1,1)) \\ &= t + \frac{(t-1)}{(t-3)} + \frac{(1-t)(t-2)}{(t-3)} = 1.\end{aligned}$$

Thus, in this particular system, none of the app users balk at the store and the overall throughput is same as the arrival rate for the app users, which is $\Lambda_T = 1$. Thus, the combined average throughput for this system, as a function of the randomization probability, is,

$$\lambda_T + \lambda_N = 1 + \frac{2(t^2 - 2t + 2)}{t^2 - 7t + 12}. \tag{B.35}$$

Using Lemma 11(i) we replace t in (B.35) by the probability, θ, that an app user orders online and obtain,

$$\lambda_T + \lambda_N = \begin{cases} \frac{4}{3} & \text{if } 0 < \frac{c_q}{c_{wT}} \leq \frac{7}{6}, \\ 1 + \frac{18c_q^2 - 36c_q c_{wT} + 20c_{wT}^2}{36c_q^2 - 51c_q c_{wT} + 18c_{wT}^2} & \text{if } \frac{7}{6} < \frac{c_q}{c_{wT}} < 2, \\ \frac{4}{3} & \text{if } 2 \leq \frac{c_q}{c_{wT}} < \infty. \end{cases}$$

■

Patient Scenario

System: Omnichannel with information

Lemma 13. *(i) The average per period consumer surplus for the app users in the omnichannel system with information for the Patient Scenario is given by,*

$$C_T = \begin{cases} v - \frac{2}{3}c_q - \frac{2}{9}c_{wT} & \text{if } 0 < \frac{c_q}{c_{wT}} < 1, \\ v - \frac{1}{3}c_q - \frac{5}{9}c_{wT} & \text{if } 1 \leq \frac{c_q}{c_{wT}} < \frac{4}{3}, \\ v - c_{wT} & \text{if } \frac{4}{3} \leq \frac{c_q}{c_{wT}} < \infty, \end{cases}$$

where $\frac{4}{3} \leq \frac{v}{c_{wT}} < 2$.

(ii) The average per period consumer surplus for the non-app users in the omnichannel system with information for the Patient Scenario is given by,

$$C_N = \begin{cases} \frac{1}{3} \cdot \left(v - \frac{2}{3}c_{wN}\right) & \text{if } 0 < \frac{c_q}{c_{wT}} < 1, \\ \frac{1}{4} \cdot \left(v - \frac{2}{3}c_{wN}\right) & \text{if } 1 \leq \frac{c_q}{c_{wT}} < \frac{4}{3}, \\ \frac{1}{4} \cdot \left(v - \frac{2}{3}c_{wN}\right) & \text{if } \frac{4}{3} \leq \frac{c_q}{c_{wT}} < \infty, \end{cases}$$

where $\frac{2}{3} \leq \frac{v}{c_{wN}} < \frac{4}{3}$.

Proof. An app user arriving at the market, upon observing queue length L, orders online if $U_o(L) > U_s(L)$ and chooses the offline option otherwise. By comparing $U_o(L)$ and $U_s(L)$, from Table B.1, we find the ordering strategy for the app users in terms of c_q and c_{wT} as summarized in Table 2.2.

Now, given the customer strategy, we next compute the steady state probabilities $\overline{\pi}_2(L)$.

If $0 < \frac{c_q}{c_{wT}} < 1$, using Table 2.2 we have

$$\mathcal{M}_{21} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{cccc} 0 & 1 & 2 & 3 \end{array} \\ \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \end{array}, \qquad \mathcal{M}_{22} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{cccc} 0 & 1 & 2 & 3 \end{array} \\ \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \end{array}.$$

The steady state probabilities are given by $\overline{\pi}_2(0) = \frac{1}{3}, \quad \overline{\pi}_2(1) = \frac{2}{3}, \quad \overline{\pi}_2(2) = 0$. Given these probabilities, the average consumer surplus for app users is given by,

$$\begin{aligned} C_T &= \overline{\pi}_2(0) \cdot (v - c_q) + \overline{\pi}_2(1) \cdot \left(v - \frac{c_q}{2} - \frac{c_{wT}}{3} \right) \\ &= v - \frac{2}{3} c_q - \frac{2}{9} c_{wT}. \end{aligned}$$

Similarly, when $1 \leq \frac{c_q}{c_{wT}} < \frac{4}{3}$, using Table 2.2 we have

$$\mathcal{M}_{21} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{cccc} 0 & 1 & 2 & 3 \end{array} \\ \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \end{array}, \qquad \mathcal{M}_{22} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{cccc} 0 & 1 & 2 & 3 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \end{array}.$$

The steady state probabilities, $\overline{\pi}_2(L)$, are given by, $\overline{\pi}_2(0) = \frac{1}{6}, \quad \overline{\pi}_2(1) = \frac{2}{3}, \quad \overline{\pi}_2(2) = \frac{1}{6}$. Given these, probabilities, the average consumer surplus for app users is given by,

$$\begin{aligned} C_T &= \overline{\pi}_2(0) \cdot (v - c_{wT}) + \overline{\pi}_2(1) \cdot \left(v - \frac{c_q}{2} - \frac{c_{wT}}{3}\right) + \overline{\pi}_2(2) \cdot (v - c_{wT}) \\ &= v - \frac{1}{3}c_q - \frac{5}{9}c_{wT}. \end{aligned}$$

Finally, when $\frac{4}{3} \leq \frac{c_q}{c_{wT}} < \infty$, using Table 2.2 we have

$$\mathcal{M}_{21} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{cccc} 0 & 1 & 2 & 3 \end{array} \\ \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \end{array}, \qquad \mathcal{M}_{22} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{cccc} 0 & 1 & 2 & 3 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \end{array}.$$

The steady state probabilities, $\overline{\pi}_2(L)$, are given by, $\overline{\pi}_2(0) = 0, \quad \overline{\pi}_2(1) = \frac{1}{2}, \quad \overline{\pi}_2(2) = \frac{1}{2}$. Given these, probabilities, the average consumer surplus for app users is given by,

$$\begin{aligned} C_T &= \overline{\pi}_2(1) \cdot (v - c_{wT}) + \overline{\pi}_2(2) \cdot (v - c_{wT}) \\ &= v - c_{wT}. \end{aligned}$$

Combining all the expressions for C_T, we have,

$$C_T = \begin{cases} v - \frac{2}{3}c_q - \frac{2}{9}c_{wT} & \text{if } 0 < \frac{c_q}{c_{wT}} < 1, \\ v - \frac{1}{3}c_q - \frac{5}{9}c_{wT} & \text{if } 1 \leq \frac{c_q}{c_{wT}} < \frac{4}{3}, \\ v - c_{wT} & \text{if } \frac{4}{3} \leq \frac{c_q}{c_{wT}} < \infty, \end{cases} \tag{B.36}$$

where, the parameters need to satisfy (B.24), i.e.,

$$\frac{4}{3} \leq \frac{v}{c_{wT}} < 2.$$

Next, we compute the average per period consumer surplus for the non-app users. Since, non-app users only join an empty system, we have,

$$C_N = \overline{\pi}_0(0) \cdot \mathbb{P}(A_t = 1) \cdot \Big(v - \frac{c_{wN}}{\mu}\Big).$$

If $0 < \frac{c_q}{c_{wT}} < 1$, the steady state probability, $\overline{\pi}_0(0)$, is given by $\overline{\pi}_0(0) = \frac{2}{3}$. Thus, we have,

$$C_N = \frac{2}{3} \cdot \frac{1}{2} \cdot \Big(v - \frac{c_{wN}}{\mu}\Big) = \frac{1}{3} \cdot \Big(v - \frac{2}{3}c_{wN}\Big).$$

If $1 \leq \frac{c_q}{c_{wT}} < \frac{4}{3}$, the steady state probability, $\overline{\pi}_0(0)$, is given by $\overline{\pi}_0(0) = \frac{1}{2}$. Thus, we have,

$$C_N = \frac{1}{2} \cdot \frac{1}{2} \cdot \Big(v - \frac{c_{wN}}{\mu}\Big) = \frac{1}{4} \cdot \Big(v - \frac{2}{3}c_{wN}\Big).$$

Finally, if $\frac{4}{3} \leq \frac{c_q}{c_{wT}} < \infty$, the steady state probability, $\overline{\pi}_0(0)$, is given by $\overline{\pi}_0(0) = \frac{1}{2}$. Thus, we have,

$$C_N = \frac{1}{2} \cdot \frac{1}{2} \cdot \Big(v - \frac{c_{wN}}{\mu}\Big) = \frac{1}{4} \cdot \Big(v - \frac{2}{3}c_{wN}\Big).$$

Combining all the expressions for C_N, the average consumer surplus for the non-app users is given by,

$$C_N = \begin{cases} \frac{1}{3} \cdot \left(v - \frac{2}{3} c_{wN}\right) & \text{if } 0 < \frac{c_q}{c_{wT}} < 1, \\ \frac{1}{4} \cdot \left(v - \frac{2}{3} c_{wN}\right) & \text{if } 1 \leq \frac{c_q}{c_{wT}} < \frac{4}{3}, \\ \frac{1}{4} \cdot \left(v - \frac{2}{3} c_{wN}\right) & \text{if } \frac{4}{3} \leq \frac{c_q}{c_{wT}} < \infty, \end{cases} \tag{B.37}$$

where, the parameters need to satisfy (B.34), i.e.,

$$\frac{2}{3} \leq \frac{v}{c_{wN}} < \frac{4}{3}.$$

■

Lemma 14. *The combined average per period throughput in the omnichannel system with information for the Patient Scenario is given by,*

$$\lambda_T + \lambda_N = \begin{cases} \frac{4}{3} & \textit{if } 0 < \frac{c_q}{c_{wT}} < 1, \\ \frac{5}{4} & \textit{if } 1 \leq \frac{c_q}{c_{wT}} < \frac{4}{3}, \\ \frac{5}{4} & \textit{if } \frac{4}{3} \leq \frac{c_q}{c_{wT}} < \infty. \end{cases}$$

Proof. If $0 < \frac{c_q}{c_{wT}} < 1$, the average throughput for non-app users is given by $\overline{\pi}_0(0) \cdot \mathbb{P}(A_t = 1) = \frac{2}{3} \cdot \frac{1}{2} = \frac{1}{3}$. From, Table 2.2, we know that if $0 < \frac{c_q}{c_{wT}} < 1$ then app users always order online. This implies that the throughput for app users is same as their arrival rate, i.e. $\lambda_o = \Lambda_T = 1$. Moreover, since all app users order online, none of the app users order at the store, i.e. $\lambda_s = 0$. Thus the combined throughput is $\lambda_N + \lambda_o + \lambda_s = \frac{4}{3}$.

If $1 \leq \frac{c_q}{c_{wT}} < \frac{4}{3}$, then the average throughput for non-app users is given by, $\overline{\pi}_0(0) \cdot \mathbb{P}(A_t = 1) = (\hat{\pi}_0(0,0) + \hat{\pi}_0(0,1)) \cdot \mathbb{P}(A_t = 1) = (\frac{1}{3} + \frac{1}{6}) \cdot \frac{1}{2} = \frac{1}{4}$. Average throughput for the app users who order at the store is given by $\Lambda_T \cdot (\hat{\pi}_1(0,1) + \hat{\pi}_1(1,1)) = 1 \cdot (\frac{1}{12} + \frac{1}{12}) = \frac{1}{6}$. We know from Table 2.2 that the app users order online only if they observe either $L = 1$ or $L = 2$. Thus, the throughput for app users who order online is given by, $\Lambda_T \cdot (\overline{\pi}_2(1) + \overline{\pi}_2(2)) = 1 \cdot (\frac{2}{3} + \frac{1}{6}) = \frac{5}{6}$. Thus, the combined throughput is $\lambda_N + \lambda_s + \lambda_o = \frac{1}{4} + \frac{1}{6} + \frac{5}{6} = \frac{5}{4}$.

Finally, if $\frac{4}{3} \leq \frac{c_q}{c_{wT}} < \infty$, then the average throughput for non-app users is given by, $\overline{\pi}_0(0) \cdot \mathbb{P}(A_t = 1) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$. Average throughput for the app users who order at the store is given by $\Lambda_T \cdot (\hat{\pi}_1(0,1) + \hat{\pi}_1(1,1)) = 1 \cdot (\frac{1}{4} + \frac{1}{4}) = \frac{1}{2}$. We know from Table 2.2 that the app users order online only if they observe $L = 2$. Thus, the throughput for app users who order online is given by, $\Lambda_T \cdot \overline{\pi}_2(2) = 1 \cdot \frac{1}{2}$. Thus, the combined throughput is $\lambda_N + \lambda_s + \lambda_o = \frac{1}{4} + \frac{1}{2} + \frac{1}{2} = \frac{5}{4}$. ■

Patient Scenario

System: Single channel

Lemma 15. *The average per period combined consumer surplus in the single channel system for the Patient Scenario is given by,*

$$C_T + C_N = \frac{4}{3}v - \frac{10}{9}c_{wT} - \frac{2}{9}c_{wN}.$$

where $\frac{4}{3} \leq \frac{v}{c_{wT}} < 2$ *and* $\frac{2}{3} \leq \frac{v}{c_{wN}} < \frac{4}{3}$.

Proof. In the Patient scenario $\tau_s = 2$, i.e. app users join only if they observe an empty system or there is a single order in the system. For this system, the probability that an app user arriving at the store observes an empty system and a system with one order, are given by $\overline{\pi}_1(0) = \frac{1}{3}$ and $\overline{\pi}_1(1) = \frac{2}{3}$ respectively. Thus the average per period consumer surplus for app users is given by,

$$\begin{aligned} C_T &= \overline{\pi}_1(0) \cdot \left(v - \frac{c_{wT}}{\mu}\right) + \overline{\pi}_1(1) \cdot \left(v - 2\frac{c_{wT}}{\mu}\right) \\ &= v - \frac{5}{3} \cdot \frac{c_{wT}}{\mu} = v - \frac{10}{9} c_{wT}. \end{aligned} \tag{B.38}$$

Next, we compute the average per period consumer surplus for non-app users. The probability that an app users arriving at the store observes an empty system is given by $\overline{\pi}_0(0) = \frac{2}{3}$. Thus we have,

$$\begin{aligned} C_N &= \mathbb{P}(A_t = 1) \cdot \overline{\pi}_0(0) \cdot \left(v - \frac{c_{wN}}{\mu}\right) \\ &= \frac{1}{2} \cdot \frac{2}{3} \cdot \left(v - \frac{2}{3} c_{wN}\right) = \frac{v}{3} - \frac{2}{9} c_{wN}. \end{aligned} \tag{B.39}$$

■

Lemma 16. *The combined average per period throughput in the single channel system for the Patient Scenario is given by,*

$$\lambda_T + \lambda_N = \frac{4}{3}.$$

Proof. In this single channel system, app users order at the store only if they observe either an empty system or a system with a single order. Thus, the average throughput

for app users is given by,

$$\lambda_T = \Lambda_T \cdot (\hat{\pi}_1(0,1) + \hat{\pi}_1(1,1)) = \frac{1}{3} + \frac{2}{3} = 1.$$

The non-app users only join an empty system. Thus, the average throughput for non-app users is given by,

$$\lambda_N = \Lambda_N \cdot \mathbb{P}(A_t = 1) \cdot (\overline{\pi}_0(0)) = \frac{1}{2} \cdot \frac{2}{3} = \frac{1}{3}.$$

■

Impatient Scenario

System: Omnichannel without information

Lemma 17. *(i) The probability, θ, that an app user arriving at the market orders online in the omnichannel system without information for the Impatient Scenario, is given by,*

$$\theta = \begin{cases} 1 & \text{if } 0 < \dfrac{c_q}{c_{wT}} \leq \dfrac{1}{6} + \dfrac{3}{4} \cdot \dfrac{v}{c_{wT}}, \\ \dfrac{12(v - c_q)}{3v - 2c_{wT}} & \text{if } \dfrac{1}{6} + \dfrac{3}{4} \cdot \dfrac{v}{c_{wT}} < \dfrac{c_q}{c_{wT}} < \dfrac{v}{c_{wT}}, \\ 0 & \text{if } \dfrac{v}{c_{wT}} \leq \dfrac{c_q}{c_{wT}} < \infty, \end{cases}$$

where $\dfrac{2}{3} \leq \dfrac{v}{c_{wT}} < \dfrac{4}{3}$.

(ii) The average per period consumer surplus for the app users in the omnichannel system

without information for the Impatient Scenario, is given by,

$$C_T = \begin{cases} v - \frac{2}{3}c_q - \frac{2}{9}c_{wT} & \text{if } 0 < \frac{c_q}{c_{wT}} \leq \frac{1}{6} + \frac{3}{4} \cdot \frac{v}{c_{wT}}, \\ \frac{v}{2} - \frac{c_{wT}}{3} & \text{if } \frac{1}{6} + \frac{3}{4} \cdot \frac{v}{c_{wT}} < \frac{c_q}{c_{wT}} < \frac{v}{c_{wT}}, \\ \frac{v}{2} - \frac{c_{wT}}{3} & \text{if } \frac{v}{c_{wT}} \leq \frac{c_q}{c_{wT}} < \infty, \end{cases}$$

where $\frac{2}{3} \leq \frac{v}{c_{wT}} < \frac{4}{3}$.

(iii) The average per period consumer surplus for the non-app users in the omnichannel system without information for the Impatient Scenario, is given by,

$$C_N = \begin{cases} \frac{1}{3} \cdot \left(v - \frac{2}{3}c_{wN}\right) & \text{if } 0 < \frac{c_q}{c_{wT}} \leq \frac{1}{6} + \frac{3}{4} \cdot \frac{v}{c_{wT}}, \\ \frac{1}{2} \cdot \frac{36c_q^2 - 45c_q v - 18c_q c_{wT} + 18v^2 + 6vc_{wT} + 4c_{wT}^2}{(3c_q - 2c_{wT})(3v - 2c_{wT})} \cdot \left(v - \frac{2}{3}c_{wN}\right) & \text{if } \frac{1}{6} + \frac{3}{4} \cdot \frac{v}{c_{wT}} < \frac{c_q}{c_{wT}} < \frac{v}{c_{wT}}, \\ \frac{1}{2} \cdot \left(v - \frac{2}{3}c_{wN}\right) & \text{if } \frac{v}{c_{wT}} \leq \frac{c_q}{c_{wT}} < \infty, \end{cases}$$

where $\frac{2}{3} \leq \frac{v}{c_{wN}} < \frac{4}{3}$ *and* $\frac{2}{3} \leq \frac{v}{c_{wT}} < \frac{4}{3}$.

Proof. The steady state probabilities of the number of orders in the system after Event 2, is given by,

$$\overline{\pi}_2(0) = \frac{t}{4-t}, \quad \overline{\pi}_2(1) = \frac{2(2-t)}{(4-t)}, \tag{B.40}$$

where t denotes the probability that an app user arriving at the market orders online. Next, we compute the expected utilities of ordering online and choosing the offline option, for an app user arriving at the market. We will have the same expected utility expressions for ordering online as in (B.21) and (B.22). Thus, we have

$$U_o(0) = v - c_q, \quad U_o(1) = v - \frac{c_q}{2} - \frac{c_{wT}}{3}.$$

Now, since in the Impatient scenario $\tau_s = \lfloor \frac{v\mu}{c_{wT}} \rfloor = 1$, we have $1 \leq \frac{v\mu}{c_{wT}} < 2$. Since, $\mu = 1 \cdot s + 2 \cdot \frac{1}{2} = \frac{3}{2}$, this implies,

$$\frac{2}{3} \leq \frac{v}{c_{wT}} < \frac{4}{3}. \tag{B.41}$$

The utility of choosing the offline option from (2.2) is given by,

$$U_s(L) = \mathbb{E}_{S_t, A_{t+1}}[\max(0, \hat{U}_s(L))]$$

where $\hat{U}_s(L)$ is given by,

$$\hat{U}_s(L) = v - ((L - S_t)^+ + n_{t+1}(L) + 1) \cdot \frac{c_{wT}}{\mu}.$$

Thus, we have,

$$U_s(0) = \mathbb{E}_{S_t, A_{t+1}}[\max(0, v - (n_{t+1}(0) + 1) \cdot \frac{c_{wT}}{\mu})].$$

We know that $n_{t+1}(L) = \min(A_{t+1}, (\tau_n - (L - S_t)^+)^+)$. Thus, $n_{t+1}(0) = \min(A_{t+1}, \tau_n)$. Since, we have assumed that $\tau_n = 1$, we have, $n_{t+1}(0) = A_{t+1}$. Using this, we can rewrite

$U_s(0)$ as,

$$U_s(0) = \mathbb{E}_{A_{t+1}}[\max(0, v - (A_{t+1}+1)\cdot \frac{c_{wT}}{\mu})].$$

Since, A_{t+1} is either 0 with probability $a = \frac{1}{2}$, or 1 with probability $1-a = \frac{1}{2}$, and (B.41) holds, we can rewrite $U_s(0)$ as,

$$U_s(0) = (1-a)\cdot\left(v - \frac{c_{wT}}{\mu}\right) = \frac{v}{2} - \frac{c_{wT}}{3}.$$

Since, $n_{t+1}(1) = \min(A_{t+1}, (\tau_n - (1-S_t)^+)^+) = A_{t+1}$, we have the exact same expression for $U_s(1)$ as that of $U_s(0)$. That implies,

$$U_s(1) = \frac{v}{2} - \frac{c_{wT}}{3}.$$

We summarize all the utility expressions for $U_o(0)$, $U_o(1)$, $U_s(0)$ and $U_s(1)$ in Table B.2.

Now, we compute the expected utilities of ordering online and choosing the offline option

L	$U_o(L)$	$U_s(L)$
0	$v - c_q$	$\frac{v}{2} - \frac{c_{wT}}{3}$
1	$v - \frac{c_q}{2} - \frac{c_{wT}}{3}$	$\frac{v}{2} - \frac{c_{wT}}{3}$

Table B.2. Utilities of ordering online and choosing the offline option by an app user when there are L order in the system

by app users using steady-state probabilities (B.40). Expected utility of ordering online is given by,

$$\overline{U}_o(t) = \frac{t(c_q - v)}{t-4} - \frac{(2t-4)(\frac{c_q}{2} - v + \frac{c_{wT}}{3})}{t-4}. \tag{B.42}$$

Since, $U_s(0) = U_s(1)$, the expected utility of choosing the offline option is given by,

$$\overline{U}_s(t) = \frac{v}{2} - \frac{c_{wT}}{3}. \tag{B.43}$$

Now, the app users randomize their choice of channel only if they are indifferent between the online and the offline option. Thus, if $0 < t < 1$, then we have, $\overline{U}_o(t) = \overline{U}_s(t)$. Equating (B.42) and (B.43) we get,

$$t = \frac{12(v - c_q)}{3v - 2c_{wT}}. \tag{B.44}$$

Applying the condition $1 > t > 0$ along with condition (B.41), we get,

$$\frac{1}{6} + \frac{3}{4} \cdot \frac{v}{c_{wT}} < \frac{c_q}{c_{wT}} < \frac{v}{c_{wT}}. \tag{B.45}$$

Finally, we characterize the average per period consumer surplus for app users as follows:

$$C_T = \begin{cases} v - \frac{2}{3}c_q - \frac{2}{9}c_{wT} & \text{if } 0 < \frac{c_q}{c_{wT}} \leq \frac{1}{6} + \frac{3}{4} \cdot \frac{v}{c_{wT}}, \\ \frac{v}{2} - \frac{c_{wT}}{3} & \text{if } \frac{1}{6} + \frac{3}{4} \cdot \frac{v}{c_{wT}} < \frac{c_q}{c_{wT}} < \frac{v}{c_{wT}}, \\ \frac{v}{2} - \frac{c_{wT}}{3} & \text{if } \frac{v}{c_{wT}} \leq \frac{c_q}{c_{wT}} < \infty, \end{cases} \tag{B.46}$$

where the parameters need to satisfy (B.41).

Now, we compute the average consumer surplus for non-app users in this system. We recall that in this system, non-app users join only if they observe an empty queue. The steady state probability is given by,

$$\overline{\pi}_0(0) = 1 - t + \frac{2t}{4 - t}$$

where t denotes the probability of an app user ordering online. Thus, the average per period consumer surplus for non-app users is given by,

$$C_N = \overline{\pi}_0(0) \cdot \mathbb{P}(A_t = 1) \cdot (v - \frac{c_{wN}}{\mu}).$$

Substituting the expression for t, as given in (B.44), in the expression for $\overline{\pi}_0(0)$, and using the condition (B.45) along with the fact that $\mu = \frac{3}{2}$ we can rewrite C_N as,

$$C_N = \begin{cases} \frac{1}{3} \cdot \left(v - \frac{2}{3}c_{wN}\right) & \text{if } 0 < \frac{c_q}{c_{wT}} \leq \frac{1}{6} + \frac{3}{4} \cdot \frac{v}{c_{wT}}, \\ \frac{1}{2} \cdot \frac{36c_q^2 - 45c_q v - 18c_q c_{wT} + 18v^2 + 6vc_{wT} + 4c_{wT}^2}{(3c_q - 2c_{wT})(3v - 2c_{wT})} \cdot \left(v - \frac{2}{3}c_{wN}\right) & \\ & \text{if } \frac{1}{6} + \frac{3}{4} \cdot \frac{v}{c_{wT}} < \frac{c_q}{c_{wT}} < \frac{v}{c_{wT}}, \\ \frac{1}{2} \cdot \left(v - \frac{2}{3}c_{wN}\right) & \text{if } \frac{v}{c_{wT}} \leq \frac{c_q}{c_{wT}} < \infty, \end{cases} \tag{B.47}$$

where the non-app users' wait-sensitivity satisfies,

$$\frac{2}{3} \leq \frac{v}{c_{wN}} < \frac{4}{3}. \tag{B.48}$$

■

Lemma 18. *The combined average per period throughput in the omnichannel system without information for the Impatient Scenario, is given by,*

$$\lambda_T + \lambda_N = \begin{cases} \frac{4}{3} & \text{if } 0 < \frac{c_q}{c_{wT}} \leq \frac{1}{6} + \frac{3}{4} \cdot \frac{v}{c_{wT}}, \\ \frac{3v - 2c_{wT}}{3c_q - 2c_{wT}} & \text{if } \frac{1}{6} + \frac{3}{4} \cdot \frac{v}{c_{wT}} < \frac{c_q}{c_{wT}} < \frac{v}{c_{wT}}, \\ 1 & \text{if } \frac{v}{c_{wT}} \leq \frac{c_q}{c_{wT}} < \infty, \end{cases}$$

where $\frac{2}{3} \leq \frac{v}{c_{wT}} < \frac{4}{3}$.

Proof. Steady state probability that a non-app user arriving at the store observes an empty system is,

$$\overline{\pi}_0(0) = 1 - t + \frac{2t}{4 - t},$$

where t denotes the probability that an app user arriving at the market orders online. Average per period throughput for non-app users is,

$$\lambda_N = \mathbb{P}(A_t = 1) \cdot \overline{\pi}_0(0) = \frac{1}{2} \cdot \frac{t^2 - 3t + 4}{4 - t}.$$

The steady state probability of an app user arriving at the store and observing an empty system is $\hat{\pi}_1(0, 1) = \frac{1}{2} - \frac{t}{2}$ where t denotes the probability of an app user ordering online. Average per period throughput for app users is,

$$\begin{aligned} \lambda_T = \lambda_o + \lambda_s &= \Lambda_T \cdot (t + \hat{\pi}_0(0, 1)) \\ &= (t + \frac{1}{2} - \frac{t}{2}) = \frac{1}{2} + \frac{t}{2}. \end{aligned}$$

Thus, the combined average throughput for this system, as a function of the probability of online ordering, t, is given by,

$$\lambda_N + \lambda_T = \frac{1}{2} \cdot \frac{t^2 - 3t + 4}{4 - t} + \frac{1}{2} + \frac{t}{2} = \frac{4}{4 - t}. \tag{B.49}$$

Using (17)(i) we replace t in (B.49) by the probability, θ, of an app user ordering online, and obtain,

$$\lambda_T + \lambda_N = \begin{cases} \frac{4}{3} & \text{if } 0 < \frac{c_q}{c_{wT}} \leq \frac{1}{6} + \frac{3}{4} \cdot \frac{v}{c_{wT}}, \\ \frac{3v - 2c_{wT}}{3c_q - 2c_{wT}} & \text{if } \frac{1}{6} + \frac{3}{4} \cdot \frac{v}{c_{wT}} < \frac{c_q}{c_{wT}} < \frac{v}{c_{wT}}, \\ 1 & \text{if } \frac{v}{c_{wT}} \leq \frac{c_q}{c_{wT}} < \infty, \end{cases}$$

where $\frac{2}{3} \leq \frac{v}{c_{wT}} < \frac{4}{3}$. ■

Impatient Scenario

System: Omnichannel with information

Lemma 19. *(i) The average per period consumer surplus for the app users in the omnichannel system with information for the Impatient Scenario is given by,*

$$C_T = \begin{cases} v - \frac{2}{3}c_q - \frac{2}{9}c_{wT} & \text{if } 0 < \frac{c_q}{c_{wT}} < \frac{1}{3} + \frac{1}{2} \cdot \frac{v}{c_{wT}}, \\ \frac{9}{10}v - \frac{2}{5}c_q - \frac{1}{3}c_{wT} & \text{if } \frac{1}{3} + \frac{1}{2} \cdot \frac{v}{c_{wT}} \leq \frac{c_q}{c_{wT}} < \frac{v}{c_{wT}}, \\ \frac{v}{2} - \frac{c_{wT}}{3} & \text{if } \frac{v}{c_{wT}} \leq \frac{c_q}{c_{wT}} < \infty, \end{cases}$$

where $\frac{2}{3} \leq \frac{v}{c_{wT}} < \frac{4}{3}$.

(ii) The average per period consumer surplus for the non-app users in the omnichannel system with information for the Impatient Scenario is given by,

$$C_N = \begin{cases} \frac{1}{3} \cdot \left(v - \frac{2}{3}c_{wN}\right) & \text{if } 0 < \frac{c_q}{c_{wT}} < \frac{1}{3} + \frac{1}{2} \cdot \frac{v}{c_{wT}}, \\ \frac{3}{10} \cdot \left(v - \frac{2}{3}c_{wN}\right) & \text{if } \frac{1}{3} + \frac{1}{2} \cdot \frac{v}{c_{wT}} \leq \frac{c_q}{c_{wT}} < \frac{v}{c_{wT}}, \\ \frac{1}{2} \cdot \left(v - \frac{2}{3}c_{wN}\right) & \text{if } \frac{v}{c_{wT}} \leq \frac{c_q}{c_{wT}} < \infty, \end{cases}$$

where, $\frac{2}{3} \leq \frac{v}{c_{wN}} < \frac{4}{3}$ *and* $\frac{2}{3} \leq \frac{v}{c_{wT}} < \frac{4}{3}$.

Proof. An app user, upon observing queue length L, orders online if $U_o(L) > U_s(L)$ and chooses the offline option otherwise. By comparing $U_o(L)$ and $U_s(L)$, from Table B.2, we find the ordering strategy for the app users in terms of v, c_q and c_{wT} as summarized in Table 2.1. Now, given the customer strategy, we next compute the steady state probabilities, $\overline{\pi}_2(L)$.

If $0 < \frac{c_q}{c_{wT}} < \frac{1}{3} + \frac{1}{2} \cdot \frac{v}{c_{wT}}$, using Table 2.1 we have

$$\mathcal{M}_{21} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{cccc} 0 & 1 & 2 & 3 \end{array} \\ \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \end{array}, \qquad \mathcal{M}_{22} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{cccc} 0 & 1 & 2 & 3 \end{array} \\ \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \end{array}.$$

The steady state probabilities are given by $\overline{\pi}_2(0) = \frac{1}{3},\ \overline{\pi}_2(1) = \frac{2}{3}$. Given these probabilities, the average consumer surplus for app users is given by,

$$\begin{aligned} C_T &= \overline{\pi}_2(0) \cdot (v - c_q) + \overline{\pi}_2(1) \cdot \left(v - \frac{c_q}{2} - \frac{c_{wT}}{3}\right) \\ &= v - \frac{2}{3}c_q - \frac{2}{9}c_{wT}. \end{aligned}$$

Similarly, when $\frac{1}{3} + \frac{1}{2} \cdot \frac{v}{c_{wT}} \leq \frac{c_q}{c_{wT}} < \frac{v}{c_{wT}}$, using Table 2.1 we have

$$\mathcal{M}_{21} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{cccc} 0 & 1 & 2 & 3 \end{array} \\ \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \end{array}, \qquad \mathcal{M}_{22} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{cccc} 0 & 1 & 2 & 3 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \end{array}.$$

The steady state probabilities, $\overline{\pi}_2(L)$, are given by, $\overline{\pi}_2(0) = \frac{1}{5},\ \overline{\pi}_2(1) = \frac{4}{5}$. Given these, probabilities, the average consumer surplus for app users is given by,

$$\begin{aligned} C_T &= \overline{\pi}_2(0) \cdot \left(\frac{v}{2} - \frac{c_{wT}}{3}\right) + \overline{\pi}_2(1) \cdot \left(v - \frac{c_q}{2} - \frac{c_{wT}}{3}\right) \\ &= \frac{9}{10}v - \frac{2}{5}c_q - \frac{1}{3}c_{wT}. \end{aligned}$$

Finally, when $\frac{v}{c_{wT}} \leq \frac{c_q}{c_{wT}} < \infty$, using Table 2.1 we have

$$\mathcal{M}_{21} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{cccc} 0 & 1 & 2 & 3 \end{array} \\ \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \end{array}, \qquad \mathcal{M}_{22} = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{cccc} 0 & 1 & 2 & 3 \end{array} \\ \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \end{array}.$$

The steady state probabilities, $\overline{\pi}_2(L)$, are given by, $\overline{\pi}_2(0) = 0, \overline{\pi}_2(1) = 1$. Given these, probabilities, the average consumer surplus for app users is given by,

$$C_T = \overline{\pi}_2(1) \cdot \left(\frac{v}{2} - \frac{c_{wT}}{3}\right) = \frac{v}{2} - \frac{c_{wT}}{3}.$$

Combining all the expressions for C_T, we have,

$$C_T = \begin{cases} v - \frac{2}{3}c_q - \frac{2}{9}c_{wT} & \text{if } 0 < \frac{c_q}{c_{wT}} < \frac{1}{3} + \frac{1}{2} \cdot \frac{v}{c_{wT}}, \\ \frac{9}{10}v - \frac{2}{5}c_q - \frac{1}{3}c_{wT} & \text{if } \frac{1}{3} + \frac{1}{2} \cdot \frac{v}{c_{wT}} \leq \frac{c_q}{c_{wT}} < \frac{v}{c_{wT}}, \\ \frac{v}{2} - \frac{c_{wT}}{3} & \text{if } \frac{v}{c_{wT}} \leq \frac{c_q}{c_{wT}} < \infty, \end{cases} \tag{B.50}$$

where, the parameters need to satisfy (B.41), i.e.,

$$\frac{2}{3} \leq \frac{v}{c_{wT}} < \frac{4}{3}.$$

Next, we compute the average per period consumer surplus for the non-app users. Since, non-app users only join an empty system, we have,

$$C_N = \overline{\pi}_0(0) \cdot \mathbb{P}(A_t = 1) \cdot \Big(v - \frac{c_{wN}}{\mu}\Big).$$

If $0 < \frac{c_q}{c_{wT}} < \frac{1}{3} + \frac{1}{2} \cdot \frac{v}{c_{wT}}$, the steady state probability, $\overline{\pi}_0(0)$, is given by $\overline{\pi}_0(0) = \frac{2}{3}$. Thus, we have,

$$C_N = \frac{2}{3} \cdot \frac{1}{2} \cdot \Big(v - \frac{c_{wN}}{\mu}\Big) = \frac{1}{3} \cdot \Big(v - \frac{2}{3}c_{wN}\Big).$$

If $\frac{1}{3} + \frac{1}{2} \cdot \frac{v}{c_{wT}} \leq \frac{c_q}{c_{wT}} < \frac{v}{c_{wT}}$, the steady state probability, $\overline{\pi}_0(0)$, is given by $\overline{\pi}_0(0) = \frac{3}{5}$. Thus, we have,

$$C_N = \frac{3}{5} \cdot \frac{1}{2} \cdot \Big(v - \frac{c_{wN}}{\mu}\Big) = \frac{3}{10} \cdot \Big(v - \frac{2}{3}c_{wN}\Big).$$

Finally, if $\frac{v}{c_{wT}} \leq \frac{c_q}{c_{wT}} < \infty$, the steady state probability, $\overline{\pi}_0(0)$, is given by $\overline{\pi}_0(0) = 1$. Thus, we have,

$$C_N = 1 \cdot \frac{1}{2} \cdot \Big(v - \frac{c_{wN}}{\mu}\Big) = \frac{1}{2} \cdot \Big(v - \frac{2}{3}c_{wN}\Big).$$

Combining all the expressions for C_N, the average consumer surplus for the non-app users is given by,

$$C_N = \begin{cases} \frac{1}{3} \cdot \Big(v - \frac{2}{3}c_{wN}\Big) & \text{if } 0 < \frac{c_q}{c_{wT}} < \frac{1}{3} + \frac{1}{2} \cdot \frac{v}{c_{wT}}, \\ \frac{3}{10} \cdot \Big(v - \frac{2}{3}c_{wN}\Big) & \text{if } \frac{1}{3} + \frac{1}{2} \cdot \frac{v}{c_{wT}} \leq \frac{c_q}{c_{wT}} < \frac{v}{c_{wT}}, \\ \frac{1}{2} \cdot \Big(v - \frac{2}{3}c_{wN}\Big) & \text{if } \frac{v}{c_{wT}} \leq \frac{c_q}{c_{wT}} < \infty, \end{cases} \tag{B.51}$$

where, the parameters need to satisfy (B.34), i.e.,

$$\frac{2}{3} \leq \frac{v}{c_{wN}} < \frac{4}{3}.$$

■

Lemma 20. *The combined average per period throughput in the omnichannel system with information for the Impatient Scenario is given by,*

$$\lambda_T + \lambda_N = \begin{cases} \frac{4}{3} & \text{if } 0 < \frac{c_q}{c_{wT}} < \frac{1}{3} + \frac{1}{2} \cdot \frac{v}{c_{wT}}, \\ \frac{6}{5} & \text{if } \frac{1}{3} + \frac{1}{2} \cdot \frac{v}{c_{wT}} \leq \frac{c_q}{c_{wT}} < \frac{v}{c_{wT}}, \\ 1 & \text{if } \frac{v}{c_{wT}} \leq \frac{c_q}{c_{wT}} < \infty, \end{cases}$$

where $\frac{2}{3} \leq \frac{v}{c_{wT}} < \frac{4}{3}$.

Proof. If $0 < \frac{c_q}{c_{wT}} < \frac{1}{3} + \frac{1}{2} \cdot \frac{v}{c_{wT}}$, the average throughput for non-app users is given by $\overline{\pi}_0(0) \cdot \mathbb{P}(A_t = 1) = \frac{2}{3} \cdot \frac{1}{2} = \frac{1}{3}$. From, Table 2.1, we know that if $0 < \frac{c_q}{c_{wT}} < \frac{1}{3} + \frac{1}{2} \cdot \frac{v}{c_{wT}}$ then app users always order online. This implies that the throughput for app users is same as their arrival rate, i.e. $\lambda_o = \Lambda_T = 1$. Moreover, since all app users order online, none of the app users order at the store, i.e. $\lambda_s = 0$. Thus the combined throughput is $\lambda_N + \lambda_o + \lambda_s = \frac{4}{3}$.

If $\frac{1}{3} + \frac{1}{2} \cdot \frac{v}{c_{wT}} \leq \frac{c_q}{c_{wT}} < \frac{v}{c_{wT}}$, then the average throughput for non-app users is given by $\overline{\pi}_0(0) \cdot \mathbb{P}(A_t = 1) = (\hat{\pi}_0(0,0) + \hat{\pi}_0(0,1)) \cdot \mathbb{P}(A_t = 1) = (\frac{2}{5} + \frac{1}{5}) \cdot \frac{1}{2} = \frac{3}{10}$. Average throughput for the app users who order at the store is given by $\Lambda_T \cdot (\hat{\pi}_1(0,1)) = 1 \cdot (\frac{1}{10}) = \frac{1}{10}$. We know from Table 2.1 that the app users order online only if they observe $L = 1$. Thus,

the throughput for app users who order online is given by, $\overline{\pi}_2(1) = \frac{4}{5}$. Thus, the combined throughput is $\lambda_N + \lambda_s + \lambda_o = \frac{3}{10} + \frac{1}{10} + \frac{4}{5} = \frac{6}{5}$.

Finally, if $\frac{v}{c_{wT}} \leq \frac{c_q}{c_{wT}} < \infty$, then the average throughput for non-app users is given by $\overline{\pi}_0(0) \cdot \mathbb{P}(A_t = 1) = 1 \cdot \frac{1}{2} = \frac{1}{2}$. Average throughput for the app users who order at the store is given by $\Lambda_T \cdot (\hat{\pi}_1(0,1)) = 1 \cdot (\frac{1}{2}) = \frac{1}{2}$. We know from Table 2.1 that the app users never order online. Thus, the combined throughput is $\lambda_N + \lambda_s + \lambda_o = \frac{1}{2} + \frac{1}{2} + 0 = 1$. ■

Impatient Scenario

System: Single channel

Lemma 21. *(i) The average per period consumer surplus for app users in the single channel system for the Impatient Scenario is given by,*

$$C_T = \frac{v}{2} - \frac{c_{wT}}{3}.$$

(ii) The average per period consumer surplus for non-app users in the single channel system for the Impatient Scenario is given by,

$$C_N = \frac{v}{2} - \frac{c_{wN}}{3}.$$

where $\frac{2}{3} \leq \frac{v}{c_{wT}} < \frac{4}{3}$ *and* $\frac{2}{3} \leq \frac{v}{c_{wN}} < \frac{4}{3}$.

Proof. In the Impatient scenario $\tau_s = 1$, i.e. app users join only if they observe an empty system. The probability that an app user arriving at the store observes an empty system is given by $\overline{\pi}_1(0) = \frac{1}{2}$. Thus the average per period consumer surplus for app users

is given by,

$$C_T = \overline{\pi}_1(0) \cdot \left(v - \frac{c_{wT}}{\mu}\right) = \frac{1}{2} \cdot \left(v - \frac{2}{3}c_{wT}\right) = \frac{v}{2} - \frac{c_{wT}}{3}. \tag{B.52}$$

Next, we compute the average per period consumer surplus for non-app users. The probability that an app users arriving at the store observes an empty system is given by $\overline{\pi}_0(0) = 1$. Thus we have,

$$C_N = \mathbb{P}(A_t = 1) \cdot \overline{\pi}_0(0) \cdot \left(v - \frac{c_{wN}}{\mu}\right) = \frac{1}{2} \cdot \left(v - \frac{2}{3}c_{wN}\right) = \frac{v}{2} - \frac{c_{wN}}{3}. \tag{B.53}$$

■

Lemma 22. *The combined average per period throughput in the single channel system for the Impatient Scenario is given by,*

$$\lambda_T + \lambda_N = 1.$$

Proof. In this single channel system, app users order at the store only if they observe an empty system. Thus, the average throughput for app users is given by,

$$\lambda_T = \Lambda_T \cdot (\hat{\pi}_1(0,1)) = \frac{1}{2}.$$

The non-app users only join an empty system. Thus, the average throughput for non-app users is given by,

$$\lambda_N = \Lambda_N \cdot \mathbb{P}(A_t = 1) \cdot (\overline{\pi}_0(0)) = 1 \cdot \frac{1}{2} \cdot 1 = \frac{1}{2}.$$

■