QUEUING SYSTEMS WITH STRATEGIC AND LEARNING CUSTOMERS

Yichen Tu

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistics and Operations Research.

Chapel Hill
2019

Approved by:

Nur Sunar

Serhan Ziya

Nilay Tanik Argon

Chuanshu Ji

Vidyadhar Kulkarni

©2019
Yichen Tu
ALL RIGHTS RESERVED

ABSTRACT

YICHEN TU: QUEUING DESIGN WHEN CUSTOMERS ARE STRATEGIC
(Under the direction of Nur Sunar and Serhan Ziya)

In many service systems customers are strategic and can make their own decisions. In particular, customers can be delay-sensitive and they will leave the system if they think the waiting time is too long. For the service provider, it is important to understand customers' behaviors and choose the appropriate system design. This dissertation consists of two research projects.

The first project studies the pooling decision when customers are strategic. It is generally accepted that operating with a combined (i.e., pooled) queue rather than separate (i.e., dedicated) queues is beneficial mainly because pooling queues reduces long-run average sojourn time. In fact, this is a well-established result in the literature when jobs cannot make decisions and servers and jobs are identical. An important corollary of this finding is that pooling queues improves social welfare in the aforementioned setting. We consider an observable multi-server queueing system which can be operated with either dedicated queues or a pooled one. Customers are delay-sensitive and they decide to join or balk based on queue length information upon arrival. In this setting, we prove that, contrary to the common understanding, pooling queues can considerably increase the long-run average sojourn time so that the pooled system results in strictly smaller social welfare (and strictly smaller consumer surplus) than the dedicated system under certain conditions. Specifically, pooling queues leads to performance loss when the arrival-rate-to-service-rate ratio and the relative benefit of service are both large. We also prove that performance loss due to pooling queues can be significant. Our numerical studies demonstrate that pooling queues can decrease the social welfare (and the consumer surplus) by more than 95%. The benefit of pooling is commonly believed to increase with the system size. In contrast to this belief, our analysis shows that when delay-sensitive customers make rational joining decisions, the magnitude of the performance loss due to pooling can strictly increase with the system size.

The second project studies the learning behavior when customers don't have full information of the service speed. We consider a single-server queueing system where customers make joining and abandonment decisions when the service rate is unknown. We study different ways in which customers process service-related information, and how these impact the performance of a service provider. Specifically, we analyze forward-looking, myopic and naive information processing behaviors by customers. Forward-looking customers learn about the service rate in a Bayesian framework by using their observations while waiting in the queue. Moreover, they make their abandonment decisions considering both belief and potential future payoffs. On the other hand, naive customers ignore the available information when they make their decisions. We prove that regardless of the way in which the information is processed by customers, a customer's optimal joining and abandonment policy is of threshold-type. There is a rich literature that establishes that forward-looking customers are detrimental to a firm in settings different than queueing. In contrast to this common understanding, we prove that for service systems, forward-looking customers are beneficial to the firm under certain conditions.

ACKNOWLEDGEMENTS

Over the past five years, I have received a lot of help and support at UNC. I am grateful to all the people who have helped me during this journey. Firstly, I want to thank my advisors, Professor Serhan Ziya and Professor Nur Sunar, for their generous guidance and support. You have also set an example of excellence as a researcher, teacher and role model. Without your continuous help and support, I will not able to finish this thesis. I also would like to thank Professor Chuanshu Ji, Professor Nilay Argon, and Professor Vidyadhar Kulkarni for serving on my thesis committee and providing great advice on my thesis. I also want to thank my fellow graduate students who help me and support me during this long journey.

Last but not least, I also want to thank my husband, Yang Hu, and my parents for their unconditional support, love and encouragement. Without them, I cannot get to this point.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

One of the fundamental questions for services that are operated by multiple servers has been whether to run the system with separated queues or a combined one. These queueing configurations are called *dedicated* and *pooled*, respectively. It is not difficult to see why pooling separate queues might be appealing: a pooled system uses the available service capacity more efficiently because under pooling no server idles as long as there are customers waiting, a possibility that exists when individual queues are kept separated. The benefit of pooling is well established in the operations management literature: when servers are identical and customers are homogeneous in their service requirements, pooling queues is proven to improve efficiency by reducing idleness and the expected waiting time in the system.

When studying the age-old question of *to pool or not to pool*, the vast majority of the literature implicitly assumed that customers are indifferent about how long they wait for service and have no say in their joining decisions. However, a common feature of many queueing systems in practice is that customers are delay-sensitive and decide whether to join a queue depending on their expected delay. Thus, it is important to analyze systems with such customers, and understand if pooling is still preferable in these systems. This is the primary objective of this paper.

The question of whether to operate a dedicated system or a pooled system is relevant in many service settings from shipping lines at the ports to voting lines in elections (The Financial Times, 2015; Hong et al., 2015; Cattani and Schmidt, 2005; Karacostas, 2018; The NYTimes, 2016). The first project studies this question by analyzing a model in which delay-sensitive customers have access to their expected delay information (e.g., through observing the queue length or by receiving delay information) and make their joining decisions based on that information. Our model is motivated by various practical settings where the service is provided for free. Two of these settings are explained below.

The first example is the design of call centers. Many organizations are grappling with the question of whether to consolidate their call centers or not (Rodriguez, 2014; Xerox, 2013; Southwest, 2012). With consolidation, calls are processed in a single large call center, rather than separate, smaller and typically region-specific call centers. In practice, the key benefit of consolidation is believed to be the efficient use of resources due to pooling, thereby improving customer satisfaction with the same or even less number of resources (Xerox, 2013). However, these anticipated benefits do not take customer behavior into account. In many call centers, callers receive queue length or expected delay information, and based on that information, they may choose not to join the system. (See Ibrahim (2018) for a literature review of such systems.)

The second example is the design of internal services in large organizations. For such organizations, there has long been a discussion on whether support services such as information technology, consulting and purchasing should be shared across different units of the organization or administered in a decentralized manner where these services are provided within each individual unit (Schmidt, 1997; Azziz, 2014; Bondarouk, 2014). Thus, in the management of internal services, the question of whether to operate a dedicated system or a pooled one is of paramount importance. Within many organizations, such as government agencies and universities, internal services are provided for free (see, e.g., page 113 of (Armbrüster, 2006) and (UAFS, 2018)), and successful implementations of such services typically rely on information sharing, which enables members of the organization to observe and identify inefficiencies such as service congestion and delays (Campbell Public Affairs Institute, 2017). Sharing support services is aimed to improve organizational efficiency by tapping into the operational benefit of pooling (Mader and Roth, 2015; U.S. Department of the Treasury, 2017). However, the design of such services also needs to account for the user behavior: if users within an organization face long delays in their service requests, they could give up solutions offered by the organization.

It is perhaps worth noting that the dedicated designs in both of these examples share a key feature: An arrival has the options of only joining her dedicated line or leaving the system.

Motivated by these practical settings, our objective is to develop and analyze a stylized formulation that centers on the following three questions: (i) How do delay-sensitive customers' rational joining decisions alter the basic calculus for the choice between pooled and dedicated systems in services with observable queue length (or delay information)? If pooling is not always preferable in such settings, what are the conditions under which the dedicated system is preferable? (ii) How large is the performance gain due to switching from one system design to another? (iii) How does the system size impact such performance

gain? We are not aware of any prior work that provides a theoretical analysis of the comparison between pooled versus dedicated queues for an observable queueing system when customers make rational joining decisions.

In the traditional pooling literature, the long-run average throughput time is a typical performance metric used to compare pooled and dedicated systems. We also use it as a performance metric. Furthermore, studying delay-sensitive self-optimizing customers allows us to analyze an additional performance metric, i.e., social welfare, which is equivalent to consumer surplus in our setting.

The second project studies the learning behavior of customers when they don't have full information about the service speed. There is growing literature on strategic behaviors of customers since the seminar work by (Naor, 1969). Some recent papers extend Naor's model by considering incomplete information of system parameters to the customers, and study the effect of information on system performance ((Cui and Veeraraghavan, 2016),(Hu et al., 2018), (Hassin and Roet-Green, 2017a), etc).

Call center has been extensively studied in operations literature since it plays an important role to interact with customers in service industry. Customers usually need to wait on hold when they call customer service. 75% of people reported they were "highly annoyed" when they could not get someone on the phone in a reasonable amount of time according to a 2015 consumer report. As a result, customers feel frustrated and abandon the service. Delay information can be announced to customers to improve their experience. A common type of delay announcements is to inform the customers about the number of people ahead periodically ((Jouini et al., 2011)). In this setting, customers are usually unknown with the system parameters. They make abandonment decisions with the evolving information from the announcements.

Ticket queue is also prevalent in service industry. In this setting, customers come to a queue and get a ticket with their order of arrival and know the number of customer in service upon arrival. Then they can go away and check later. There is no physical queue. Customers can balk upon arrival or abandon the service later ((Pender and Jennings, 2015)). Recently some technology solutions emerge to manage the ticket queue via mobile device, kiosk or web browser (for instance, Nemo-Q and QLess), which makes real-time information communication possible. In the ticket queue literature, it is common to assume that (i) customers have constant patience, (ii) customers treat all tickets ahead as real customers and (iii) customers do not update their decisions, yet only (ii) holds ((Kuzu, 2015)). (Kuzu et al., 2017) empirically study the customers' abandonment behaviors in a ticket queue incorporating the dynamic updating decision of customers. In this model, the decision epoch follows Poisson process. The customer updates his forecast

waiting time based on her current ticket position and number of servers, and reneges if it is larger than reneging patience, which follows exponential distribution.

Motivated by these practical settings, we raise the following questions. What policy should the customer use to make joining decision and abandonment decision without full information? Is it always beneficial to be forward-looking? From the service provider's perspective, should he share the information with the customers?

We study these research problems in a discrete time single-server queue setting. Customers do not know the true service rate and have a Bernoulli prior belief on the service rate distribution. The service provider inform the customers with the queue position upon arrival. The customers make joining decisions either based on their expected payoff by joining in a myopic way, or in a more complicated way by dynamic programming. We also consider the abandonment behaviors when customers update their beliefs in a Bayesian framework based on the service completion information provided by the service provider. We characterize the structural properties of three alternative policies, i.e., simple policy, myopic policy and forward-looking policy. It is shown that being complicated may result in loss of total expected benefit for the customer. We also conduct throughput comparison between systems with different types of customers.

Our paper has three main contributions. First, to our best knowledge, there is no prior theoretical work studying customers' alternative policies in a queueing system without assuming the service parameters as common knowledge. Second, the paper provides novel insights about the realizations of each policy from customers' perspective and shows that it is not always beneficial to be bayesian and forward-looking. Finally, the throughput comparison between systems with different types of customers provides managerial insights for the service provider.

CHAPTER 2

Pooled or Dedicated Queues when Customers are Delay-Sensitive

In this chapter, we consider a system which can be run as either the dedicated system or the pooled system. We consider the optimal system choice in terms of long-run average throughput time and social welfare.

2.0.1 Summary of Main Results

Considering delay-sensitive customers' rational joining decisions in the comparison of pooled versus dedicated queues gives rise to the following three unexpected results for the observable systems.

First, Smith and Whitt (1981) establish that if every arrival joins the system (without making decisions), pooling queues is beneficial in the case of identical servers and jobs. In contrast, our paper proves (in Theorem 1-(a)) that if arriving customers decide to join or balk, the dedicated system can outperform the pooled system depending on the following two factors: (i) relative benefit of service, which is the ratio of service benefit to customer's waiting cost per unit time, and (ii) potential system load, which is the ratio of arrival rate to service rate. Specifically, if both the relative benefit of service and the potential system load are large, pooling queues strictly increases the average sojourn time and this increase is so large that, compared to the dedicated system, the pooled system results in strictly smaller social welfare.

Second, our analysis and numerical studies show that the performance improvement due to separating queues can be drastic. Specifically, our paper proves (in Theorem 2) that the percentage increase in the social welfare with dedicated queues can be arbitrarily large, compared to the case with a pooled queue.

Third, in the case of nonstrategic and identical jobs and servers, the benefit of pooling queues is well-known to increase with the number of servers (keeping the arrival rate to service rate ratio the same) (Calabrese, 1992; Benjaafar, 1995). In contrast, our paper proves (in Theorem 3) that when customers

make their own joining decisions, the magnitude of the performance loss due to pooling can strictly increase with the number of servers (keeping the arrival rate to service rate ratio the same).

To provide a complete picture, our paper also identifies conditions under which the pooled system results in smaller average sojourn time and hence larger social welfare than the dedicated system. (Those conditions can be found in Theorem 1-(b).)

Our paper also studies variants of the base model. Some of the key messages from this additional analysis (in Section 2.3) are as follows: (i) The observability of queue length or real-time expected delay information is necessary for our unexpected results to hold. (ii) The dedicated system may outperform the pooled system even when customers are allowed to choose the shortest queue in the dedicated system. (iii) All of our results extend when customers incur a fixed fee upon service completion. [1] (iv) When a social planner could charge a different service fee under each queue configuration to maximize social welfare, pooling queues improves the social welfare. Thus, the welfare advantage of pooling queues can be recovered if the social planner has the pricing lever.

2.0.2 Relevant Literature

Our paper belongs to the literature that studies pooled versus dedicated queues. To the best of our knowledge, there is no prior work that theoretically analyzes the comparison of pooled versus dedicated queues in an observable system when delay-sensitive customers make their own joining decisions. Our paper provides such an analysis.

The analysis of pooling queues has long been an interest in the queueing literature. To our knowledge, (Smith and Whitt, 1981) were the first to provide a mathematical investigation of pooling queues. (Smith and Whitt, 1981) showed that when jobs (e.g., customers) are homogeneous in their service requirements and servers are identical, pooling separate queues increases the system efficiency by reducing the expected steady-state waiting time. Since the publication of this seminal work, many articles studied the benefit of pooling queues in different contexts and under a variety of conditions. (Calabrese, 1992) provided an alternative proof to show the benefit of pooling for system efficiency. (Benjaafar, 1995) determined bounds on performance improvements through pooling. Gans et al. (2003a) illustrated the benefits of pooling call centers (in different geographical locations) into one. Using approximation formulas for a two-server queueing system, van Dijk and van der Sluis (2008) made the observation that when customers

[1] A fixed service fee represents a price-taker service provider, which emerges in a perfectly competitive service environment.

are identical, a pooled system results in smaller long-run average waiting time than its dedicated counterpart. (Andradóttir et al., 2017) showed that even if servers are subject to failures, pooling queues always results in smaller expected steady-state number of jobs in the system and hence smaller long-run average waiting average time, compared to the system with dedicated queues.

Unlike what has been established in this literature, our paper proves that when delay-sensitive customers make their joining decisions in an observable system, pooling queues can result in much worse performance than a dedicated system even with identical servers and homogeneous customers.

There have been observations that pooling parallel queues is not always beneficial and may result in performance degradation; these observations are attributed to three main factors explained in (a) through (c) below. Our paper identifies a different factor not previously identified in the pooling literature: observable queue and customers' ability to make a joining/balking decision. We now explain the aforementioned three factors in (a) through (c) below, and discuss the relevant literature.

(a) If jobs are heterogeneous in their service requirements or servers are not identical, the pooled system may perform worse than the dedicated system. (Smith and Whitt, 1981) included a numerical example with heterogeneous servers to make this point. (Rothkopf and Rech, 1987) discussed that if jobs require different service times, combining separate queues into a single one can increase the average delay. Section 5.3 of Mandelbaum and Reiman (1998) briefly discussed this effect of heterogeneous servers in a parallel multi-server setting without providing proofs (as there are no exact formulas available in that setting). Using approximation formulas for queueing models, van Dijk and van der Sluis (2008) and van Dijk and van der Sluis (2009) constructed numerical examples to illustrate the aforementioned effect of these factors.

(b) Pooling queues may also result in worse performance (e.g., larger expected steady-state waiting time) due to server slowdown and other server-related issues. (Rothkopf and Rech, 1987) argued that when service times increase due to combining separate queues into a single one, the pooled system may result in larger average delay than the dedicated one. (Gilbert and Weng, 1998) studied a setting where there are two self-interested servers and a principle that compensates servers based on their performance. In this setting, authors established that pooling queues can be undesirable for the principle due to server incentives. (Shunko et al., 2018) conducted controlled lab experiments to find an evidence of server slowdown due to pooling queues. (Jouini et al., 2008) numerically demonstrated that the dedicated system can outperform the pooled system if each agent works slower in the pooled system potentially due to decreased customer

ownership. (Song et al., 2015) empirically investigated the effects of pooling in an emergency department and found that the dedicated system is superior to the pooled system with respect to the average waiting time and the average length-of-stay. The paper attributes this to physicians' increased ownership of the patients under the dedicated system. Do et al. (2015) theoretically analyzed the implications of server slowdown due to pooling, and showed that the pooled system can result in larger expected waiting time than the dedicated system. Using a data set from a supermarket, Wang and Zhou (2017) provided an empirical evidence that pooling queues can increase the service time. The main driver of this finding was explained to be the social loafing effect with a pooled queue. (Armony et al., 2017) considered a two-server queueing system where servers can choose their long-run average service rates, and incur a cost for the expected workload or busyness. Armony et al. (2017) showed that if servers are workload-averse, pooling queues always achieves lower expected queue length but can result in larger expected work-in-process (WIP).

(c) Apart from two factors explained in (a) and (b), (Rothkopf and Rech, 1987) conjectured that when jockeying (i.e., switching from one queue to another) among parallel queues is possible for customers, under very mild conditions, the average waiting time under the pooled system can be larger than that under the dedicated system.

It is worth emphasizing that none of the papers mentioned in (a) through (c) theoretically analyzes customers that can make their own joining decisions. Unlike all of the papers mentioned above, our work provides a theoretical analysis of such self-optimizing customers in the context of pooling queues. In our problem formulation, to avoid any performance advantage to the dedicated system and to analyze the effect of customers' joining decisions in isolation, we will exclude the above factors that were previously observed to cause pooling to potentially perform worse than the dedicated system.

Lu et al. (2013) empirically analyzed a data set from a supermarket's checkout line. Considering a specific queue setting, Lu et al. (2013) found evidence that the queue length can be an important driver of customers' purchasing behaviors. Based on this, Lu et al. (2013) argued that if there existed a practical setting in which customers' purchasing behaviors under pooled and dedicated systems were both the same as the one identified by the authors, pooling queues may decrease average waiting time due to balking. Lu et al. (2013) considered a specific queue setting in their study, and hence did not empirically investigate the trade-offs between pooled and dedicated systems. (Thus, empirically identifying the aforementioned hypothetical practical setting in which customers' purchasing behaviors under pooled and dedicated systems are the same is still an open question.) Unlike Lu et al. (2013), our paper theoretically compares pooled versus dedicated

systems by considering rational customers' joining decisions. Moreover, the main performance metric in our paper is social welfare, which is the same as consumer surplus in our setting.

Our work is also relevant to the literature that studies delay-sensitive rational customers making their own decisions in observable queueing systems. (The comparison of pooled versus dedicated queues has not been investigated in this literature.) Our formulation of customers builds on the framework developed and analyzed in the seminal work by (Naor, 1969). (Naor, 1969) considered a single-server queue where customers can observe the queue length and decide whether to join the queue or balk depending on their expected net benefit of joining the queue. In his setting, a balking customer gains zero expected net benefit, while each joining customer incurs a constant waiting cost per unit time spent in the system, and receives a reward upon service completion. One of the main findings of (Naor, 1969) is that allowing customers to make their own decisions results in social welfare loss compared to the maximum achievable welfare. Many articles extend the model analyzed in (Naor, 1969) in various dimensions. The comprehensive review of these papers can be found in (Hassin and Haviv, 2003) and (Hassin, 2016a). But, two of them are especially worth highlighting here. (Debo and Veeraraghavan, 2014) extended Naor's model by carrying out an equilibrium analysis for a queue with incomplete information. In their setting, customers observe the queue length but they do not exactly know the service value and the expected service time before making a join or balk decision. The authors proved that customers' joining probability does not necessarily decrease with the queue length. (Cui and Veeraraghavan, 2016) built on Naor's model to analyze a setting where customers have different beliefs about the service time, and the service provider can reveal service information. They established that revealing service information can significantly hurt the social welfare and consumer surplus.

2.0.3 Outline of the Paper

The remainder of our paper is organized as follows. Section 3.1 introduces the model and includes preliminary analysis. Section 3.3 includes the main results and their interpretations. Section 2.3 studies several extensions of the base model. Section 3.7 includes concluding remarks. Proofs of all formal results as well as supplementary results and their proofs are presented in Appendices A.1 through A.15 of the Electronic Companion.

2.1 Model

Consider a first-come-first-served (FCFS) queueing system with $N \geq 2$ servers. The service time of each server is exponentially distributed with rate $\mu > 0$.[2] The system can be run with either dedicated queues or a pooled queue. These two alternatives will be called *dedicated* and *pooled systems*, and indexed by $j = d$ and $j = p$, respectively.

The dedicated system contains N separate queues, each served by a separate server. In this setting, a server together with its queue is called a dedicated *sub-system*. In the dedicated system, customers arrive to each queue according to a Poisson process with rate $\Lambda_d = \lambda$, and a server provides service only to customers in his own queue.[3] In contrast, in the pooled system, separate queues are combined into a single one, and customers arrive to the queue according to a Poisson process with rate $\Lambda_p = N\lambda$. Whenever a server completes serving a customer, he serves the next customer waiting in the queue. Here, Λ_d and Λ_p can be interpreted as *potential arrival rate* for a queue in the associated system. In light of this, the ratio

$$\rho \doteq \lambda/\mu \tag{2.1}$$

is called the *potential system load*. As will be explained later, the actual arrival rate to a queue is different than the potential arrival rate because the former is determined by customers' joining decisions.

Customers make their own joining decisions. Regardless of the system type, upon arrival, each customer first observes the queue length and then decides whether to join the queue or balk. If an arriving customer decides to join the queue, the customer incurs cost $c > 0$ per unit time she spends in the system. A customer gains a benefit R after service completion, and the service is free of charge. Considering a queueing system that provides free of charge service is common in the literature. Although their research questions are very different than ours, several studies analyze such systems. (See, for instance, Hassin (1985), Armony et al. (2009), Gai et al. (2011) and Haviv and Oz (2016).) Section 2.3.1 explains that all of our results and their proofs extend in a straightforward fashion if customers pay a fixed fee $f > 0$ upon service completion. In

[2] Our model considers identical servers to tease out the effect of customers' joining decisions; heterogeneous servers were already observed to cause pooling to potentially perform worse than the dedicated system.

[3] Considering a dedicated arrival stream for each server is common in the formulation of dedicated queueing systems. See, for instance, Smith and Whitt (1981) and Yu et al. (2015). Section 2.3.3 demonstrates that if customers are allowed to choose the shortest queue in the dedicated system, the key phenomenon proved in Theorem 1-(a) extends under certain conditions.

our formulation, all model parameters are common knowledge. This implies that for customers, observing the queue length is the same as observing their real-time expected sojourn time.

As in Naor (1969), if an arriving customer decides to balk, she neither gets a benefit nor incurs a cost, and hence she gains zero expected net benefit. If a customer arrives to a particular queue, the customer receives the following expected net benefit by joining the queue:

$$\mathbb{E}\left[U(n;j)\right] = R - \bar{W}_j(n+1)c, \quad j \in \{d,p\}. \tag{2.2}$$

Here, $\bar{W}_j(n+1)$ represents the expected time spent by the arriving customer in the system; for the pooled system, n represents the number of customers that are already in the system, and for the dedicated system, n corresponds to the number of customers that are already in the arrived sub-system. A customer joins the queue if and only if her expected net benefit is non-negative, which is equivalent to the following by (2.2):

$$\mathbb{E}\left[U(n;j)\right] = R - \bar{W}_j(n+1)c \geq 0;$$

otherwise she balks. This suggests that an arriving customer optimally joins the queue if and only if the number of customers in the queue and its associated service is smaller than a threshold that depends on the system type; otherwise, the customer balks.

The aforementioned optimal threshold rule implies two key characteristics of the systems in our analysis. First, the rate at which customers join the queue, which is represented by $\lambda_{e,j}$, is always smaller than the potential arrival rate Λ_j for $j \in \{d,p\}$. Second, regardless of the value of the potential system load ρ, both pooled and dedicated systems are stable.[4]

Our primary goal is to analyze the implications of pooling for *social welfare*. In doing so, we will also study the implications of pooling for long-run average time spent in the system, that is, *average sojourn time*.

Denote by W_j, the average sojourn time in the system $j \in \{d,p\}$. In our setting, the social welfare equals the consumer surplus, which is the sum of long-run average net gains of all customers in a system. As a result, the social welfare in the system $j \in \{d,p\}$ is equal to the multiplication of these two factors: (i) a single customer's long-run average net benefit $R - W_j c$ and (ii) the long-run average number of customers

[4]Similarly, the $M/M/1$ system studied by (Naor, 1969) is stable regardless of ρ. This property was further explained by (Gilboa-Freedman et al., 2014).

served, that is, *throughput*, θ_j:

$$SW_j = (R - W_j c)\theta_j \doteq \begin{cases} (R - W_j c)\lambda_{e,j} = R\lambda_{e,j} - cL_j & \text{if } j = p, \\ (R - W_j c)\lambda_{e,j} N = R\lambda_{e,j} N - cL_j N & \text{if } j = d. \end{cases} \tag{2.3}$$

Here, L_p is the long-run average number of customers in the pooled system, L_d represents its counterpart in *one* of the N dedicated sub-systems, and the throughput θ_j satisfies the following:

$$\theta_j = \begin{cases} \lambda_{e,j} & \text{if } j = p, \\ \lambda_{e,j} N & \text{if } j = d. \end{cases} \tag{2.4}$$

According to (2.3), the average sojourn time and the throughput are the two key determinants of the social welfare.

2.1.1 Preliminary Analysis

To avoid trivialities, this paper focuses on a case where

$$k \doteq \left\lfloor \frac{R\mu}{c} \right\rfloor \geq 1. \tag{2.5}$$

Here, $\lfloor \cdot \rfloor$ is the standard floor function. The condition (2.5) means that a customer always joins an empty system.

Lemma 1. *In the pooled system, an arriving customer joins the queue if and only if the number of customers already in the system is* $n \leq K - 1$*, where*

$$K \doteq \left\lfloor \frac{RN\mu}{c} \right\rfloor. \tag{2.6}$$

Furthermore, in the pooled system, for $K > N$, the average sojourn time and social welfare are respectively given by

$$W_p = \frac{\sum_{i=0}^{N-1} \frac{N^i}{i!} i\rho^i + \frac{N^N}{N!} \sum_{i=N}^{K} i\rho^i}{\left(\sum_{i=0}^{N-1} \frac{N^i}{i!} \rho^i + \frac{N^N}{N!} \sum_{i=N}^{K-1} \rho^i\right) N\lambda}, \tag{2.7}$$

$$SW_p = \left(1 - \frac{\frac{N^N}{N!}\rho^K}{\sum_{i=0}^{N-1} \frac{N^i}{i!}\rho^i + \frac{N^N}{N!}\sum_{i=N}^{K} \rho^i}\right) RN\lambda - \frac{\sum_{i=0}^{N-1} \frac{N^i}{i!} i\rho^i + \frac{N^N}{N!} \sum_{i=N}^{K} i\rho^i}{\sum_{i=0}^{N-1} \frac{N^i}{i!}\rho^i + \frac{N^N}{N!}\sum_{i=N}^{K} \rho^i} c, \tag{2.8}$$

where ρ is as defined in (2.1).

Lemma 2. *In the dedicated system, an arriving customer joins the queue if and only if the number of customers already in that sub-system is $n \leq k-1$, where k is as defined in (2.5). Furthermore, in the dedicated system, the average sojourn time and social welfare are respectively given by*

$$W_d = \frac{\sum_{i=0}^{k} i\rho^i}{\left(\sum_{i=0}^{k-1} \rho^i\right)\lambda} \quad \textit{and} \quad SW_d = \left(1 - \frac{\rho^k}{\sum_{i=0}^{k} \rho^i}\right) RN\lambda - \frac{\sum_{i=0}^{k} i\rho^i}{\sum_{i=0}^{k} \rho^i} Nc, \tag{2.9}$$

where ρ is as defined in (2.1).

Based on Lemmas 1 and 2, hereafter, we refer to k as the *balking threshold* in the dedicated system and K as the *balking threshold* in the pooled system. Note that the balking thresholds satisfy

$$K \geq Nk. \tag{2.10}$$

2.2 Analysis

We begin the analysis with one of our main results.

Theorem 1. *There exist constants η and $\bar{\eta}$ such that the following results hold:*

(a) *The dedicated system results in* **(i)** *strictly smaller average sojourn time and* **(ii)** *strictly larger social welfare than the pooled system, i.e., $W_d < W_p$ and $SW_d > SW_p$, respectively, if*

$$\rho > 1 \quad \textit{and} \quad R/c > \eta, \tag{2.11}$$

where η is finite when $\rho > 1$, and does not depend on either R or c.

(b) *The pooled system results in* **(i)** *smaller average sojourn time and* **(ii)** *strictly larger social welfare than*

the dedicated system, i.e., $W_p \leq W_d$ and $SW_p > SW_d$, respectively, if either

$$(\star)\ \ R/c < (N+1)/(N\mu) \quad \textit{or} \quad (\star\star)\ \ \rho < 1 \quad \textit{and} \quad R/c > \bar{\eta}, \tag{2.12}$$

where $\bar{\eta}$ is finite when $\rho < 1$, and does not depend on either R or c.

Theorem 1-(a) establishes that when customers make their own joining decisions, pooling queues can be undesirable even with identical servers and customers. There are two main drivers of this result: (i) The system is observable. Each customer makes the joining decision based on her *own* expected sojourn time, and joins as long as this sojourn time is (weakly) smaller than the longest one, which is experienced by the customer who joins at the balking threshold. Thus, customers do not account for negative externality they impose on customers behind in their joining decisions, leading to very congested systems. This situation is in contrast to an unobservable queueing system where customers make their joining decisions based on the average sojourn time across all customers that join the system. (ii) When the system has a large potential load, i.e., $\rho > 1$, the stationary probability of having l customers in the system is convex increasing in l.

Theorem 1-(b) proves that the benefit of pooling is recovered under conditions (2.12). A detailed explanation of Theorem 1 is provided in Section 2.2.1.

2.2.1 Explanation of Theorem 1

We will first provide a step-by-step analysis to establish and explain Theorem 1-(a), which is our key result. Because the throughput is an important determinant of social welfare by (2.3), we begin our analysis with the following result.

Proposition 1. *The dedicated system results in strictly smaller throughput than the pooled system, i.e., $\theta_d < \theta_p$.*

The rationale behind Proposition 1 is as follows. The dedicated system has N sub-systems, and each of them is a single-server queueing system with a balking threshold k. Thus, there can be a situation where a customer arrives to a dedicated sub-system, and finds out that there are already k customers in the sub-system whereas other dedicated sub-systems have not reached their balking thresholds. Such a situation is not possible in the pooled system because the pooled system has a single queue with balking threshold K, which is more than N times the dedicated system balking threshold k by (2.10). In addition,

the pooled system also reduces idleness; this makes it less likely for the pooled system to operate at the balking threshold than the dedicated system. Because of all these reasons, the pooled system results in smaller balking probability and hence larger throughput than the dedicated system, as proved in Proposition 1.

Proposition 1 and (2.3) suggest that even if the dedicated system has a smaller average sojourn time than the pooled system, the pooled system can still outperform the dedicated one in terms of social welfare. The dedicated system can outperform the pooled system in terms of social welfare only when the former has a sufficiently lower average sojourn time that offsets the lower throughput.

We now introduce a "*scaled queueing system*," i.e., *SQ system*, as a bridge for the comparison between the dedicated and pooled systems. We consider the SQ system because comparing the dedicated system with the SQ system or comparing the pooled system with the SQ system is analytically more tractable than directly comparing the dedicated system with the pooled system.

Definition 2.1. An *SQ system* is a single-server queueing system indexed by $j = s$ with the following properties: **(a)** Customers arrive to the system according to a Poisson process with rate λN. **(b)** The service time has an exponential distribution with rate μN. **(c)** Each arriving customer balks if and only if the number of customers already in the system is larger than K (as defined in (2.6)); otherwise, she joins the system.

Let W_s and SW_s be the average sojourn time and social welfare in the SQ system, respectively. Then, we have the following result.

Proposition 2. (**SQ versus Pooled**) *Compared to the pooled system, the SQ system results in* **(a)** *strictly smaller average sojourn time, i.e.,* $W_s < W_p$*, and* **(b)** *strictly larger social welfare, i.e.,* $SW_s > SW_p$.

There are two key observations related to the SQ system: (i) The SQ and pooled systems have the same balking threshold K. (ii) The service rate in the SQ system is larger than the one in the pooled system for any given number of customers in the system, and the former is strictly larger than the latter when the number of customers in the system is small. Then, by (i) and (ii), the SQ system results in strictly smaller average sojourn time than the pooled system, as proved in Proposition 2-(a). The throughput in the SQ system is strictly larger than the one in the pooled system by the proof of Proposition 2. [5][6] This and Proposition 2-(a) immediately imply Proposition 2-(b).

[5]See inequality (A.41) in the proof of Proposition 2.

[6]This result and Proposition 1 imply $\theta_s > \theta_p > \theta_d$ where θ_s is the throughput in the SQ system.

Proposition 3. (**SQ versus Dedicated**) *The dedicated system results in* **(a)** *strictly smaller average sojourn time and* **(b)** *strictly larger social welfare than the SQ system, i.e.,* $W_d < W_s$ *and* $SW_d > SW_s$*, respectively, if*

$$\rho > 1 \quad \textit{and} \quad R/c > \eta, \tag{2.13}$$

where η *is the same constant as the one in Theorem 1-*(a).

Remark 2.2.1. There exists a constant $\widetilde{\eta}$ such that if $\rho < 1$ and $R/c > \widetilde{\eta}$, then $W_d > W_s$ and $SW_d < SW_s$. This result is stated as Proposition 20 in Appendix A.5 of the Electronic Companion. Its proof can be found in the same appendix.

Let us explain the rationale behind Proposition 3. To that end, denote by L_s the steady-state average number of customers in the SQ system. Recall that θ_s and θ_d are the throughputs of the SQ and dedicated systems, respectively, and L_d is the steady-state average number of customers in each of the N dedicated sub-systems. Based on this notation, let us first explain part (a) of the proposition. As will be explained in detail below, $L_s > NL_d$ when $\rho > 1$. Consequently, if (2.13) holds, then $W_s > W_d$ due to the following two reasons: (i) When the relative benefit of service is large, i.e., $R/c > \eta$, the balking thresholds in each dedicated sub-system and the SQ system are both large, which implies that the throughputs θ_s and θ_d are very close to each other. (ii) By Little's Law, $W_s = L_s/\theta_s$ and $W_d = NL_d/\theta_d$. Since θ_s and θ_d are very close to each other, this implies part (a), i.e., $W_s > W_d$. To explain part (b) of the proposition, we note the following. As stated above, when $R/c > \eta$, the throughputs θ_s and θ_d are very close to each other, and thus, the average sojourn time is the determining factor in the comparison of social welfare in the SQ and dedicated systems. As a result, part (b) holds.

We now explain why $L_s > NL_d$ when $\rho > 1$. For this purpose, let $\pi_d(i)$ (respectively, $\pi_s(i)$) be the steady-state probability of having i customers in a dedicated sub-system (respectively, in the SQ system). Then, $L_d = \sum_{i=0}^{k} \pi_d(i) i$ and $L_s = \sum_{i=0}^{K} \pi_s(i) i$, where k and K are the balking thresholds in a dedicated sub-system and the SQ system, respectively. When $\rho > 1$, the steady-state probabilities $\pi_d(i)$ and $\pi_s(i)$ are convex increasing in i, the number of customers. Thus, in the summations $L_d = \sum_{i=0}^{k} \pi_d(i) i$ and $L_s = \sum_{i=0}^{K} \pi_s(i) i$, the weight of the term i is convex increasing in i when $\rho > 1$. Because these weights are probabilities that sum up to 1, we note that when $\rho > 1$, a larger i has a larger weight in both L_s and L_d. Since $K > k$, the support of $\pi_s(i)$, namely $\{i : i = 0, 1, \ldots, K\}$, extends to larger values of i, than the

support of $\pi_d(i)$, which is $\{i : i = 0, 1, \ldots, k\}$. As a result, $\pi_s(\cdot)$ puts even more weight to larger values of i, compared to $\pi_d(\cdot)$. Therefore, when $\rho > 1$, the convex increasing property of the steady-state distribution is more pronounced in the SQ system than in the dedicated system. Combining this with (2.10), we deduce that the sum $L_s = \sum_{i=0}^{K} \pi_s(i)i$ is strictly larger than the sum $NL_d = N \sum_{i=0}^{k} \pi_d(i)i$.

By Propositions 2 and 3, if (2.13) holds, $W_d < W_s < W_p$ and $SW_d > SW_s > SW_p$. Thus, we have Theorem 1-(a). These orderings are in contrast to the classical understanding that is based on no-customer-balking assumption. In the absence of customer balking, one would have $W_s < W_p < W_d$ and $SW_s > SW_p > SW_d$, where $j = s$ here is the modified scaled system that satisfies properties (a) and (b) in Definition 2.1, and assumes no balking.[7]

We now explain the conditions in Theorem 1-(b). If (2.12)-($\star$) holds, a joining customer immediately enters the service in both dedicated and pooled systems because $k = 1$ and $K = N$ under that condition. Thus, dedicated and pooled systems have the same average sojourn time W_j and provide the same long-run average net benefit to each joining customer. This and strictly larger throughput in the pooled system (by Proposition 1) imply strictly larger social welfare for the pooled system if (2.12)-($\star$) holds.

The conditions in (2.12)-($\star\star$) can be explained as follows. When the benefit is large (i.e., $R/c > \bar{\eta}$), balking thresholds are large in both systems. With a relatively small potential load, i.e., $\rho < 1$, dedicated and pooled systems barely achieve their balking thresholds, implying very small expected number of balking customers for both systems. Thus, the pooled and dedicated systems have very close throughputs under (2.12)-($\star\star$). Moreover, under these conditions, there is significant idleness in the dedicated system. As a result, pooling results in smaller average sojourn time by reducing idleness in the system. This and Proposition 1 imply larger social welfare for the pooled system.

2.2.2 Numerical Comparison of the Pooled and Dedicated Systems

Figure 2.1 pictures conditions under which the dedicated system outperforms the pooled system for a numerical example. In this figure, $SW_d > SW_p$ if and only if $(R/c, \rho)$ pair lies in either Region III or Region IV (i.e., the region to the right of the solid line). Thus, for a wide range of parameters, the

[7] When customers cannot balk, throughputs are the same for $j \in \{s, d, p\}$ because every arrival joins. As the service rate is larger in the modified scaled system than in the pooled system for any given number of customers in the system, $W_s < W_p$ in the absence of customer balking. (The proof of this statement is similar to the proof of Proposition 2, and hence omitted). We already know from Smith and Whitt (1981) that $W_p < W_d$ when customers cannot balk. Combining these, we have $W_s < W_p < W_d$. Then, $SW_s > SW_p > SW_d$ in the absence of customer balking because $SW_j = N\lambda(R - cW_j)$ for $j \in \{s, d, p\}$ in that setting.

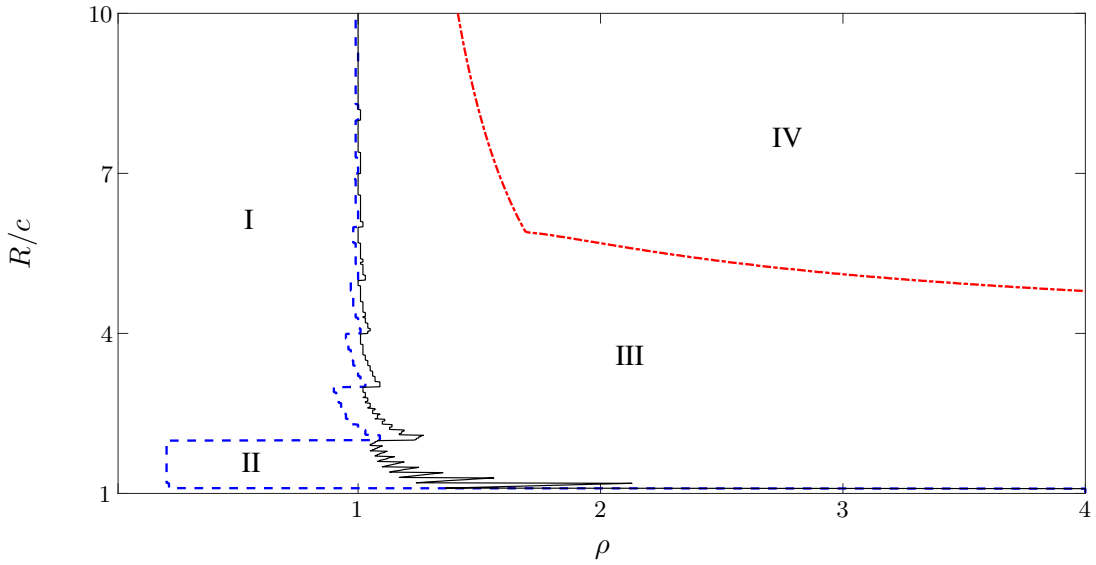

Figure 2.1: Comparison of pooled and dedicated systems when $c = 1$, $\mu = 1$, $N = 10$. The displayed dashed and solid boundaries between regions are non-smooth and zigzagged because of the floor function in k and K. In this figure, $W_d \geq W_p$ if and only if $(R/c, \rho)$ pair is in Region I, and $SW_p \geq SW_d$ if and only if $(R/c, \rho)$ pair is either in Region I or in Region II. Region IV is as explained in Section 3.2.

dedicated system results in strictly larger social welfare than the pooled system. Region IV corresponds to the parameter space identified in (2.11). We can see that the sufficient condition (2.11) constitutes a large portion of the parameter set in which $SW_d > SW_p$.

Figure 2.1 demonstrates that the dedicated system results in strictly larger social welfare than the pooled system for a given service rate if and only if R/c is not too small and $\rho > \rho_{SW}$ for some constant ρ_{SW}. In this figure, $W_d < W_p$ if and only if $(R/c, \rho)$ pair lies in either Region II, III or IV. This suggests that the dedicated system results in strictly smaller average sojourn time than the pooled system if and only if R/c is not too small and $\rho > \rho_W$ for some constant $\rho_W \leq \rho_{SW}$. We have $\rho_{SW} \geq \rho_W$ because if the dedicated system results in larger social welfare than the pooled system at a given ρ, then, by (2.3) and Proposition 1, it must also result in strictly smaller average sojourn time than the pooled system at the same ρ.

Observe from Figure 2.1 that $\rho_W < 1$ for certain values of R/c. Our further numerical analysis shows that ρ_{SW} can also be smaller than 1 for some R/c. This means that $\rho > 1$ is not a necessary condition for the superior performance of the dedicated system. We also numerically observed that $SW_d > SW_p$ for $\rho < 1$ when R/c is not too large and $R\mu/c$ is not an integer. Thus, we conjecture that the parameter region in which $SW_d > SW_p$ with $\rho < 1$ is much smaller compared to the one with $\rho > 1$.

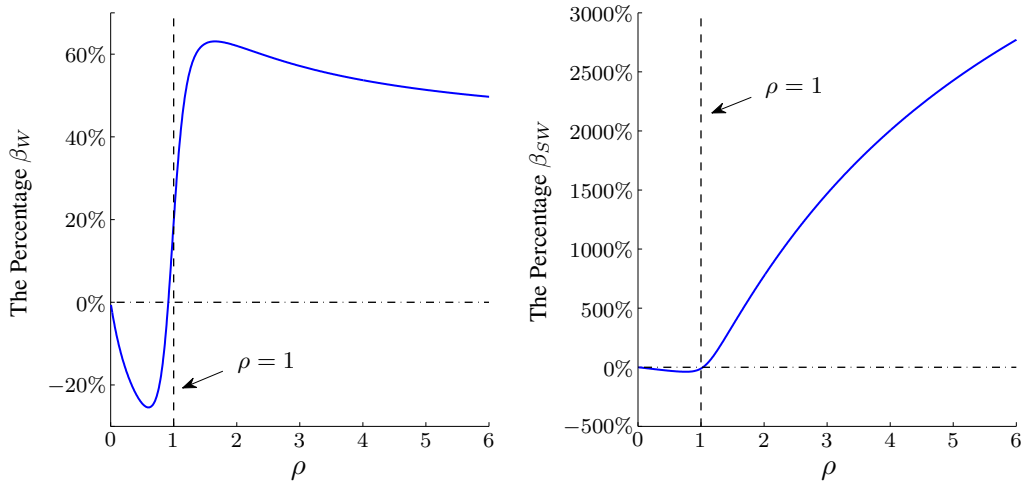

Figure 2.2: Throughput time ratio and social welfare ratio. The following parameters are used: $R = 75$, $c = 4$, $N = 10$ and $\mu = 0.15$.

Note from (2.5) and (2.6) that the potential arrival rate λ does not affect the balking thresholds k and K. However, Theorem 1 and Figure 2.1 demonstrate that λ plays an important role in the comparison between the dedicated system and the pooled system through ρ. Figure 2.2 above sheds more light on the effect of ρ on the comparison between the pooled and dedicated systems.

Figure 2.2 demonstrates the percentages $\beta_W \doteq (W_p - W_d)/W_d \times 100\%$ and $\beta_{SW} \doteq (SW_d - SW_p)/SW_p \times 100\%$ for a numerical example. A key message from this figure is that the dedicated system can result in significantly larger social welfare than the pooled system for large ρ.[8] Among other properties, the steep increase in β_W around $\rho = 1$ in Figure 2.2 shows that when ρ is close to 1, the average sojourn time can increase in the potential load significantly faster under the pooled system, than under the dedicated system. This increase eventually leads to welfare loss under pooling. Note from Figure 2.2 that the explained sojourn time phenomenon cannot be observed as $\rho \to 0$ or $\rho \to \infty$. (The aforementioned sojourn time observations are analytically verified by (A.73), (A.75) and (A.76) in Lemma 13, which is in Appendix A.6 of the Electronic Companion.) Overall, Figure 2.2 underscores the importance of judiciously evaluating the pooled and dedicated systems for services, as the relative performance of a system can be very sensitive to a change in ρ.

[8]One can show that the ratio SW_d/SW_p can increase unboundedly as $\rho \to \infty$ for example when $R\mu/c = k + 1/N$. Our further numerical studies suggest that the ratio SW_d/SW_p can be bounded as $\rho \to \infty$ when $R\mu/c$ is an integer.

2.2.3 The Impact of Benefit R and Number of Servers N

Theorem 1 naturally brings forth the following question: What is the percentage *improvement* in the social welfare SW if the dedicated system is implemented instead of the pooled system? Theorem 2 below answers this question by identifying a lower bound for the achievable aforementioned percentage under certain conditions.

The parameter R is one of the determinants of SW because it affects both the throughput and the average sojourn time through the balking thresholds k and K. In Theorem 2, we will include R as an argument of $SW(\cdot)$ to emphasize its dependence on R. After presenting Theorem 2, we will further discuss the effect of R.

Theorem 2. *Compared to the pooled system, the percentage increase in the social welfare under the dedicated system satisfies the following for $\rho > 1$:*

$$\max_{R} \left\{ \beta_{SW}(R) \doteq \frac{SW_d(R) - SW_p(R)}{SW_p(R)} \times 100\% \right\} > (N-2) \times 100\%. \tag{2.14}$$

Figure 2.3 displays how $\beta_{SW}(R)$ (defined in (2.14)) changes with R for a given system size. Because welfare loss under the pooled system is due to the large increase in the average sojourn time with pooling, it could also be worthwhile to see the percentage *increase* in the average sojourn time W due to pooling queues. Thus, Figure 2.3 also displays $\beta_W(R) \doteq \frac{W_p(R) - W_d(R)}{W_d(R)} \times 100\%$ with respect to R. There are a few key observations in this figure: (i) Operating a system with dedicated queues rather than a pooled one can significantly improve the social welfare in both small-scale and large-scale systems. (ii) Such significant performance gain does not require a very large R. For example, when $N = 50$, the percentage improvement in social welfare under the dedicated system is larger than 100% for $R \geq 7.5$. The reason is that pooling queues can drastically increase the average sojourn time even at moderate benefit R; in fact, the maximum $\beta_W(R)$ is typically observed at moderate R, as displayed in Figure 2.3.

Consider a sequence of systems indexed by $n = \{2, 3, \ldots\}$ such that in the n^{th} system, there are $N = n$ servers and the total potential arrival rate in the system is $n\lambda$. In this context, n can be seen as a proxy for the *system size*. Theorem 2 suggests that pooling queues can be detrimental especially when the system

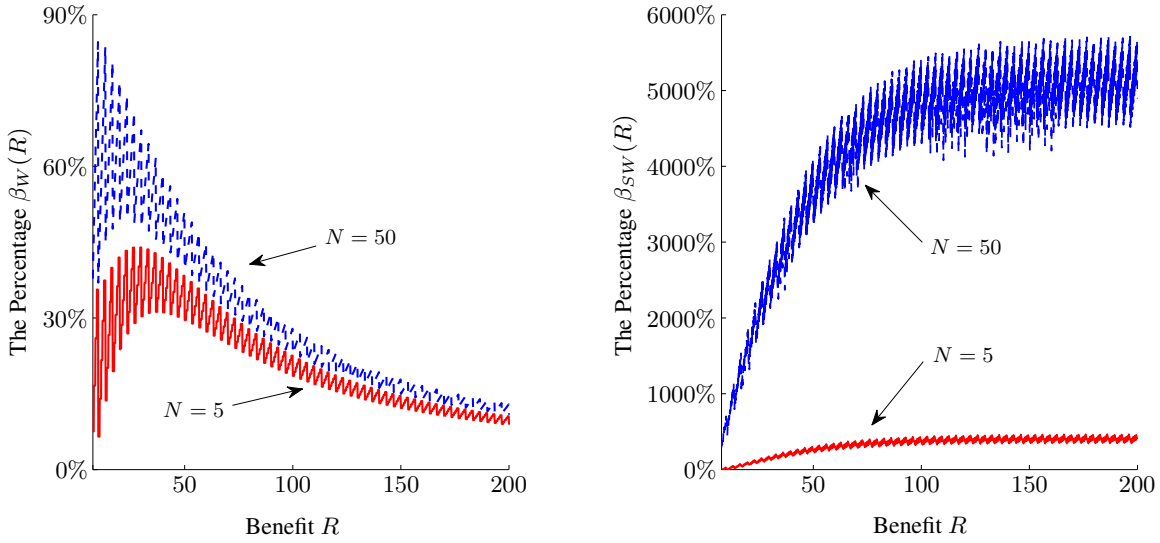

Figure 2.3: The percentages $\beta_W(R)$ and $\beta_{SW}(R)$ when $\lambda = 0.35$, $c = 1$, $\mu = 0.3$. The displayed functions are non-smooth and zigzagged because of the floor function in k and K. Total potential arrival rate in the system with N servers is λN. Thus, N can be seen as the scale of the system.

size is large. Specifically, Theorem 2 shows that, compared to the pooled system, the dedicated system may improve social welfare in a way that the percentage improvement in social welfare eventually takes values larger than any fixed value as the system size increases.

It is well established in the literature that pooling queues in a larger system provides larger performance benefits. For instance, Benjaafar (1995) demonstrates that when there is no balking, the average delay decreases with the system size when multiple $M/M/1$ systems are combined and run as a pooled system. An important implication of this observation in their setting is that the social welfare benefit of pooling also increases with the system size. In contrast, Figure 2.3 shows a numerical example where pooling in a larger system results in *larger* percentage *loss* in social welfare for each R when customers make their own joining decisions. Specifically, for any R, $\beta_{SW}(R)$ with $N = 50$ is larger than that with $N = 5$ in Figure 2.3. An important driver of this is that under the considered parameters, the percentage increase in the average sojourn time due to pooling is larger in a larger system. (See $\beta_W(R)$ in Figure 2.3.) Motivated by these observations, Figure 2.4 provides a deeper numerical analysis on the impact of the system size on the percentages $\beta_W(N)$ and $\beta_{SW}(N)$. Here, we include N as an argument of the performance metric under consideration to emphasize its dependence on N.

There are two key observations related to Figure 2.4: First and most important, the percentages $\beta_W(N)$ and $\beta_{SW}(N)$ display an increasing trend in N. Theorem 3 below will formalize this observation. Second,

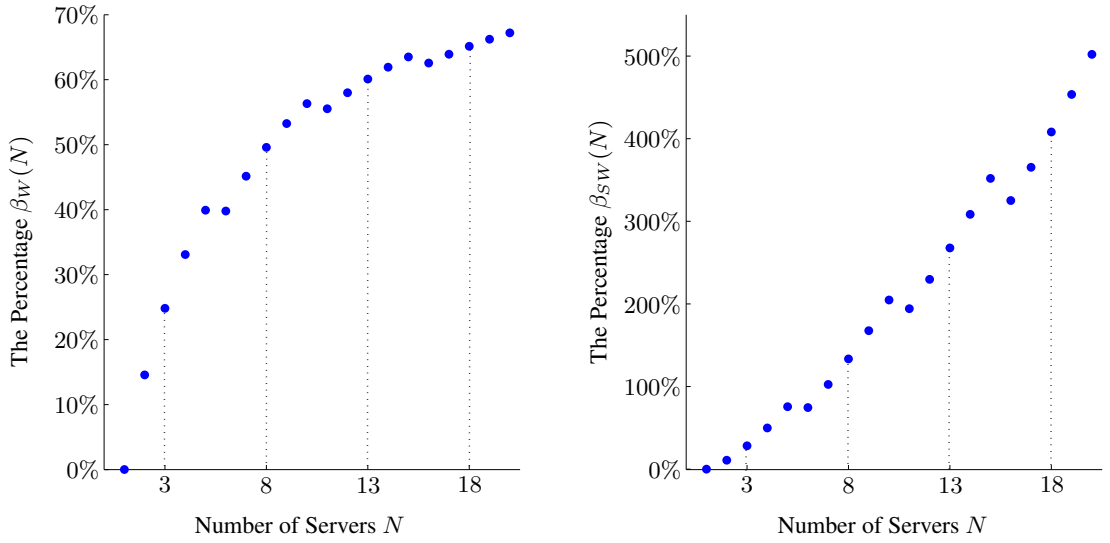

Figure 2.4: The percentages $\beta_W(N) \doteq \frac{W_p(N)-W_d(N)}{W_d(N)} \times 100\%$ and $\beta_{SW}(N) \doteq \frac{SW_d(N)-SW_p(N)}{SW_p(N)} \times 100\%$ when $\lambda = 0.35$, $c = 1$, $\mu = 0.3$.

both $\beta_W(N)$ and $\beta_{SW}(N)$ display a certain non-monotone pattern. In particular, the percentages increase with N for five data points, and switch to a different level, and then increases with N for 5 more data points. This pattern repeats itself. Our additional numerical analysis suggests that such a pattern is observed when $R\mu/c$ is not an integer (because of the floor function in k and K); if $R\mu/c$ is an integer, $\beta_W(N)$ and $\beta_{SW}(N)$ both strictly increase with N. Theorem 3 below will shed more light on this behavior.

To state Theorem 3, we shall introduce new notation. Let m be the minimum system size n that makes $Rn\mu/c$ an integer:

$$m \doteq \min\left\{n \in \mathbb{N}_+ : \frac{Rn\mu}{c} \in \mathbb{N}_+\right\}. \tag{2.15}$$

For instance, in Figure 2.4, $m = 5$. Because of this, we observe a repeating pattern (in which percentages increase with N) for every fifth consecutive data point starting from any single point on Figure 2.4. We now define a subsequence that considers every i^{th} system size in each repeating pattern. Specifically, for each $i = 1, 2, \ldots, m$, define a system size subsequence $\mathcal{S}_i \doteq \{N_{i,0}, N_{i,1}, N_{i,2}, N_{i,3}, \ldots\}$ such that

$$N_{i,\ell} = i + \ell m, \quad \ell = 0, 1, \ldots \tag{2.16}$$

For example, in Figure 2.4, the system size subsequence $\mathcal{S}_3 = \{3, 8, 13, 18, \dots\}$ includes the third system size ($i = 3$) in each repeating pattern that consists of 5 data points ($m = 5$).

Theorem 3. *Let $i = 1, 2 \dots, m$. Then, we have the following results:*

(a) *There exists a constant η_1 such that the subsequence $\beta_W(N_{i,\cdot}) \doteq \left\{ \frac{(W_p(N_{i,\ell}) - W_d(N_{i,\ell}))}{W_d(N_{i,\ell})} \times 100\%,\ \ell = 0, 1, \dots \right\}$ is non-negative and strictly increasing in the system size if $\rho > 1$ and $R/c > \eta_1$. The constant η_1 is finite when $\rho > 1$, and does not depend on either R or c.*

(b) *There exists a constant η_2 such that the subsequence $\beta_{SW}(N_{i,\cdot}) \doteq \left\{ \frac{(SW_d(N_{i,\ell}) - SW_p(N_{i,\ell}))}{SW_p(N_{i,\ell})} \times 100\%,\ \ell = 0, 1, \dots \right\}$ is non-negative and strictly increasing in the system size if $\rho > 1$ and $R/c > \eta_2$. The constant η_2 is finite when $\rho > 1$, and does not depend on either R or c.*

Figure 2.4 displays an example where both $\beta_W(N)$ and $\beta_{SW}(N)$ are positive and strictly increase in N for $N \in \mathcal{S}_3 = \{3, 8, 13, 18, \dots\}$, as suggested by Theorem 3.

There are two main drivers of Theorem 3: (i) In both dedicated sub-system and pooled system, the stationary probability of having l customers in the system is convex and increasing in l for $\rho > 1$, and (ii) the balking threshold in the pooled system increases in N. Because of these, when there is a performance loss due to pooling, the loss is exacerbated even more with an increase in the system size.

The reason for the contrast between Theorem 3 and the classical finding in Benjaafar (1995) about the effect of system size on the benefit of pooling is the following. In Benjaafar (1995), customers are not delay sensitive and join the system regardless. Therefore, that study assumes an infinite queue capacity and $\rho < 1$ for stability. In contrast, our paper considers rational joining decisions of delay-sensitive customers (which imply a finite balking threshold) and allows for $\rho > 1$.

2.3 Extensions

With this section, we aim to check the robustness of our key result in Theorem 1-(a). This analysis will also help us further investigate what causes the dedicated system to outperform the pooled system in terms of social welfare.

2.3.1 Fixed Price

Consider a setting where all modeling elements are the same as in Section 3.1, except each customer pays a fixed price $f > 0$ upon service completion. This setting represents a service environment where the

service provider does not have any control over price. All our proofs can be extended in a straightforward fashion by replacing R with $R - f$ in k and K. Thus, Theorems 1(a)-(i) and 1-(b) hold after replacing R with $R - f$. The statement of Theorem 1-(a)-(ii) requires an additional condition that $f < \bar{f}$ for some threshold fee $\bar{f}$.

2.3.2 Optimal Pricing

Suppose that each customer pays a fee upon service completion in the system $j \in \{d, p\}$, and this service fee is set to either maximize the service provider's revenue or the social welfare. All other modeling elements are the same as in Section 3.1.

Based on this, we will compare the pooled and dedicated systems under the following two formulations:
(i) *Welfare-maximizing fee*: For each system $j \in \{d, p\}$, a service fee f_j is set to maximize the social welfare:

$$\max_{f_j \geq 0} \; SW_j \doteq (R - cW_j(f_j))\theta_j(f_j), \quad j \in \{d, p\}, \tag{2.17}$$

where $\theta_j(\cdot)$ is the throughput in the system $j \in \{d, p\}$. The social welfare does not include the term $f_j\theta_j(f_j)$ because the total collected fee is just a transfer between customers and the fee collector.
(ii) *Revenue-maximizing fee*: For each system $j \in \{d, p\}$, a service fee f_j is set to maximize the service provider's revenue:

$$\max_{f_j \geq 0} \; RV_j \doteq f_j\theta_j(f_j), \quad j \in \{d, p\}. \tag{2.18}$$

Proposition 4. **(a)** *Under formulation* (2.17)*, the maximum social welfare in the pooled system is larger than that in the dedicated system.*
(b) *Under formulation* (2.18)*, the maximum revenue in the pooled system is larger than that in the dedicated system.*

Under formulation (2.17), the service fee affects the social welfare only through the resulting balking threshold. By setting the fee, the social planner prevents the system from becoming too congested, and hence customers cannot over-utilize the system. In that case, the pooled system improves the system efficiency by reducing idleness in the system. To sum up, under formulation (2.17), in the pooled system, by setting the fee, the social planner not only changes customers' joining behaviors in a socially-optimal way but also

achieves the system efficiency. As a result, as proved in Proposition 4-(a), the pooled system outperforms the dedicated system when formulation (2.17) is considered.

Proposition 4-(b) follows from two facts: (i) For any fixed service fee, the pooled system results in strictly larger throughput than the dedicated system. The rationale behind this fact is the same as the one explained for Proposition 1. This fact implies that at any given fee, the revenue under the pooled system is strictly larger than that under the dedicated system. (ii) Under formulation (2.18), the optimal fee for the dedicated system is feasible but not necessarily optimal for the pooled system.

Lemma 3. *Under formulation* (2.18)*, the social welfare under the pooled system is strictly smaller than the one under the dedicated system if* R/c *is in a moderate range, i.e.,* $R/c \in (\underline{r}, \bar{r})$ *for some constants* $\underline{r}$ *and* $\bar{r}$*.*

Under the stated conditions in Lemma 3, the balking thresholds of dedicated and pooled systems are both equal to N, which implies that the average sojourn time under the dedicated system is strictly larger than the one under the pooled system. However, when $R/c \in (\underline{r}, \bar{r})$, the dedicated system results in such a larger throughput than the pooled system that we obtain the result in Lemma 3. For example, when R is in $(118, 173)$, $c = 7$, $N = 2$, $\lambda = 0.1$ and $\mu = 0.5$, the dedicated system results in strictly larger social welfare than the pooled system under formulation (2.18).

Apart from these results, one can show that when the relative benefit of service is moderate, the dedicated system can result in strictly smaller average sojourn time than the pooled system under formulations (2.17) and (2.18). Our additional numerical studies suggest that, under formulation (2.17), pooling queues can significantly increase the average sojourn time.

2.3.3 Join-the-Shortest-Queue Policy

In Section 3.1, there is a separate arrival stream for each queue in the dedicated system. Consider an alternative dedicated queueing system where an arriving customer observes the number of customers in each of N queues, and then decides whether to join a queue or balk. We will refer to this system as the *alternative system*, and denote it by the index $j = a$. In the alternative system, if an arriving customer decides to join, she optimally chooses the shortest queue. When multiple queues are in a tie, we assume that the customer chooses the queue with the smallest index. (Another way to break the tie is to pick a queue randomly with equal probability. This alternative tie-breaking rule would not alter any of our insights.) In

this setting, an arriving customer optimally balks if and only if each of the N dedicated sub-systems already has k customers.

It is well known in the literature that the exact analysis of the *join-the-shortest-queue (JSQ) policy* in a first-come first-served queueing system is typically intractable (Gupta et al., 2007). There were some efforts to analyze a system under the JSQ policy with 2 servers (see, for instance, (Kingman, 1961) and (Flatto and McKean, 1977)). However, even such analysis was found to be difficult (Selen et al., 2016; Haight, 1958). Because the exact analysis of the JSQ policy with more than two servers still remains infeasible (Gupta et al., 2007), the vast majority of the literature focuses on approximations or numerical analysis to evaluate the performance of the JSQ policy (see, for instance, (Grassmann, 1980), (Rao and Posner, 1987) and (Nelson and Philips, 1993), among others). In short, the general theoretical analysis of the JSQ policy with delay-sensitive rational customers is hard. However, as we explain below, under certain conditions, we can compare the performance of the alternative system with those of the pooled and dedicated systems.

The following proposition considers a queueing system with $N \geq 2$ servers.

Proposition 5. *Suppose that $K - Nk > 0$, which holds when $R\mu/c - \lfloor R\mu/c \rfloor \geq 1/N$. Then, as $\lambda \to \infty$, the average sojourn time under the alternative system is strictly smaller than the one under the pooled system, i.e., $W_a(\lambda) < W_p(\lambda)$, and the social welfare under the alternative system is strictly larger than the one under the pooled system, i.e., $SW_a(\lambda) > SW_p(\lambda)$.*

Thus, even if customers are allowed to choose which queue to join in the dedicated system, separating queues yields a superior performance under some conditions. The rationale behind Proposition 5 is as follows. As the arrival rate gets very large (i.e., $\lambda \to \infty$), both alternative and pooled systems mostly operate at their respective balking thresholds. This means that in the limit (i.e., $\lambda \to \infty$), as soon as a customer is served, a new customer joins the system. That is, as $\lambda \to \infty$, every joined customer joins as the last customer before the system reaches its balking threshold, and hence experiences the longest (feasible) expected sojourn time in the system almost surely. The condition $K - Nk > 0$ implies that the longest expected sojourn time in the pooled system is strictly larger than that in the alternative system. Thus, the average sojourn time in the pooled system is strictly larger than that in the alternative system. Furthermore, because servers are busy with probability 1 as $\lambda \to \infty$, throughputs of the pooled and alternative systems are the same and both equal to the total service rate ($N\mu$). Combining this and the aforementioned average

sojourn time comparison, we conclude that the social welfare under the pooled system is strictly smaller than that under the alternative system.

Figure 2.5 pictures a numerical example to compare the performances of alternative and pooled systems. All numerical examples in this section use exact balance equations to identify the steady-state queue length distribution in the alternative system. The left panel of Figure 2.5 displays β_W^a, which represents the percentage increase in the average sojourn time due to pooling queues. The right panel of Figure 2.5 pictures β_{SW}^a, which represents the percentage increase in the social welfare under the alternative system, compared to the case under the pooled system. Note from Figure 2.5 that the alternative system can significantly outperform the pooled system in terms of social welfare. Another observation is that $\rho > 1$ is not necessary for the alternative system to outperform the pooled system.

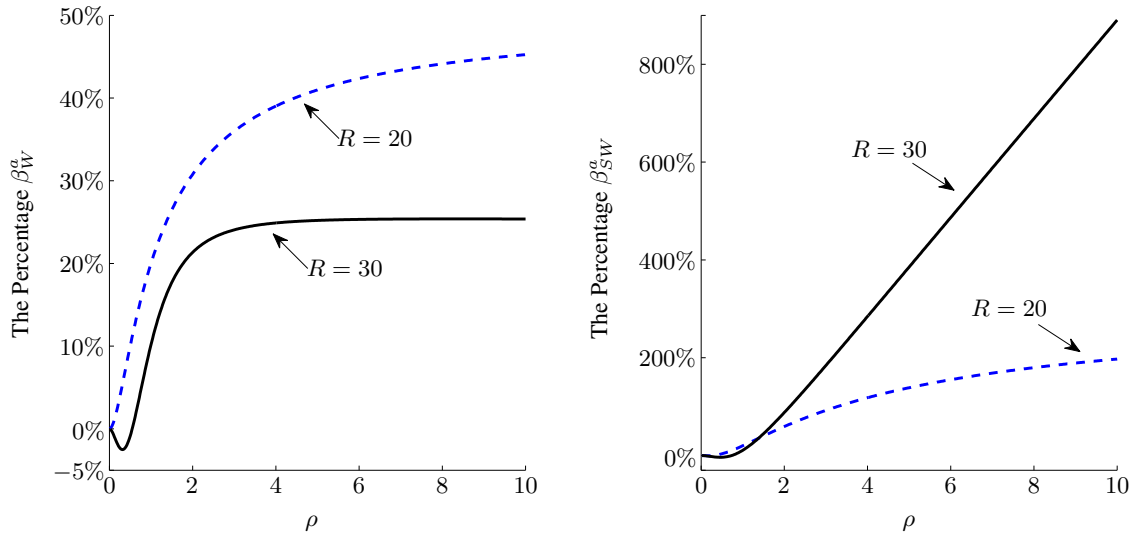

Figure 2.5: The percentages are defined as $\beta_W^a \doteq \frac{W_p - W_a}{W_a} \times 100\%$ and $\beta_{SW}^a \doteq \frac{SW_a - SW_p}{SW_p} \times 100\%$. The following parameters are used: $c = 6$, $\mu = 0.5$, $N = 2$.

Similar to the observations in Section 2.2.2, our numerical examples with $K - Nk > 0$ demonstrate that for a given service rate, the alternative system results in (i) strictly smaller average sojourn time than the pooled one if and only if $\rho > \rho_W^a$, and (ii) strictly larger social welfare than the pooled system if and only if $\rho > \rho_{SW}^a$. Table 2.1 displays these thresholds and their counterparts ρ_W and ρ_{SW} in Section 2.2.2 for a particular set of parameters. Observe that for each listed R, each of ρ_W^a and ρ_{SW}^a is strictly smaller than its counterpart in Section 2.2.2. This means that for these values of R, the parameter region in which the pooled system is dominated is larger when the pooled system is compared with the alternative system rather

Table 2.1: The thresholds with the following parameters: $\mu = 1$, $c = 2$, and $N = 2$.

R	ρ_W^a	ρ_W	ρ_{SW}^a	ρ_{SW}
5	0.55	0.7	0.74	1.04
7	0.75	0.86	0.85	1.05
9	0.84	0.92	0.89	1.04
11	0.88	0.95	0.92	1.03
13	0.9	0.97	0.93	1.03

than with the dedicated system. Our further numerical study suggests that this conclusion is valid even for large values of R.

It could be valuable to identify which system among the three (i.e., alternative, dedicated and pooled systems) maximizes social welfare. Our numerical study suggests that the alternative system can outperform both dedicated and pooled systems. For example, consider $R = 5$, $\mu = 1$, $c = 2$ and $N = 2$. Then, among the three systems, the alternative system results in maximum social welfare for medium range of ρ, i.e., when $\rho \in (\rho_{SW}^a, \bar{\rho}_{SW}^a)$ for some threshold $\bar{\rho}_{SW}^a$.[9] For that example, among the three systems, the pooled system generates the maximum social welfare for small ρ, i.e., $\rho < \rho_{SW}^a$, and the dedicated system results in maximum social welfare for large ρ, i.e., $\rho > \bar{\rho}_{SW}^a$.[10] A similar pattern is observed for all R listed in Table 2.1, as well as in many other numerical examples with $K - Nk > 0$.

2.3.4 Partial Pooling

This section considers partial pooling as an alternative system design, and demonstrates that even with the partial pooling option, the dedicated system yields the best performance for a large set of parameters. In our setting, partial pooling refers to combining only some of the separate queues (instead of all queues) to form a single line.

[9] Under this condition, the alternative system achieves the maximum social welfare among the three systems intuitively because of the following two reasons: (i) Balking probability in the pooled system is smaller than the one in the alternative system. Thus, when the potential system load is considerably large, the pooled system is more over-utilized and yields larger average sojourn time than the alternative system. The resulting increase in the average sojourn time under the pooled system is so large that the alternative system performs strictly better than the pooled system in terms of social welfare. (ii) Compared to the dedicated system, the alternative system improves the throughput by giving customers more discretion in their joining decisions. When the potential system load is moderate, the larger throughput in the alternative system translates into the larger social welfare for the alternative system.

[10] When the potential system load is large, the alternative system is much more congested than the dedicated system because customers join the former system more. As a result, under the aforementioned condition, compared to the dedicated system, the alternative system results in much larger average sojourn time, leading to strictly smaller social welfare in the alternative system.

Figure 2.6 demonstrates a numerical example where there are $N = 20$ servers and partial pooling is allowed. To focus on reasonable number of partial pooling scenarios, the example considers symmetric partial pooling, meaning that each pooled sub-system (within the partially-pooled system) contains the same number of servers. Thus, for this example, partial pooling refers to combining every $M \in \mathcal{D} \doteq \{2, 4, 5, 10\}$ queues of the $N = 20$ queues into a separate single queue. For instance, if $M = 2$, the system consists of 10 pooled sub-systems, each formed by combining 2 separate queues into one. Note that there are four alternative partial pooling designs because other than 1 and 20, $\{2, 4, 5, 10\}$ are all divisors of 20.

Figure 2.6 displays the system that results in the minimum average sojourn time and maximum social welfare among all system designs. Based on this, when R/c is not too small, there exist a threshold ρ^p_{SW} such that the dedicated system generates maximum social welfare if and only if $\rho > \rho^p_{SW}$. This implies a very large region in which the dedicated system achieves the best performance among all systems. Moreover, the parameter region in which the dedicated system achieves the maximum social welfare is a subset of the one in which the dedicated system results in minimum average sojourn time. The explained observations are similar to the ones in Section 3.3. Our further numerical study shows that the aforementioned threshold is smaller in the absence of partial pooling option, which corresponds to the setting in Section 3.1. The reason is as follows. Compared to the setting in Section 3.1, with partial pooling, there are more systems to be compared with each other. The dedicated system is less likely to perform the best among more options. Thus, the ρ threshold above which the dedicated system outperforms all other systems is smaller in the absence of the partial pooling option.

It is feasible to identify some sufficient conditions under which the dedicated system achieves the best performance among all designs. Specifically, the dedicated system outperforms all other designs under the conditions in Theorem 1-(a), except that the threshold η in that theorem should be replaced with another one. The proof of Theorem 1-(a) extends to this setting by replacing N with M when the dedicated system is compared with a partially-pooled system that consists of N/M pooled sub-systems, each with M servers.

The reason why our key result - the superiority of the dedicated system under some conditions - extends to the partial pooling setting can be explained as follows. Intuitively, when a system is partially pooled, there is no interaction between distinct sub-systems of pooled queues, and thus, each sub-system of pooled queues can be viewed as an independent system of pooled queues. Consequently, the comparison between a partially-pooled system and its dedicated counterpart is equivalent to the comparison between each pooled sub-system and the dedicated counterpart of that sub-system. To be more precise, the dedicated system

with N servers outperforms a partially-pooled system with N servers and N/M symmetric sub-systems of pooled queues if and only if the dedicated system with M servers outperforms a completely pooled system with M servers. In fact, using similar logic, it is feasible to identify similar sufficient conditions under which the dedicated system performs the best in any type of partial pooling setting (including the non-symmetric ones).

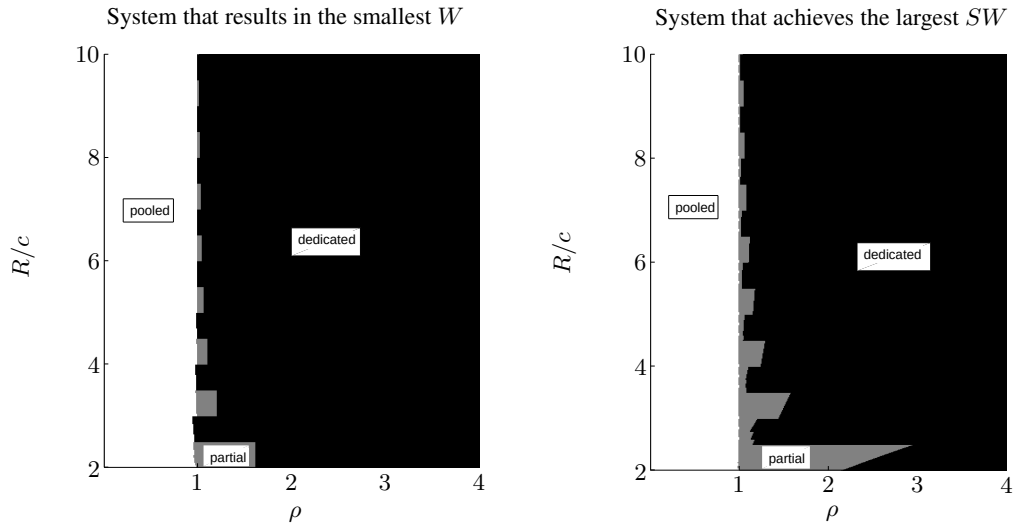

Figure 2.6: The dedicated, partially-pooled and pooled systems achieve the best performance in black, grey and white regions respectively. The following parameters are used: $c = 1$, $\mu = 1$ and $N = 20$. The displayed boundaries between regions are non-smooth because of the floor function in k and K.

As displayed in Figure 2.6, partial pooling outperforms both the dedicated and pooled systems when R/c is moderate, and ρ is strictly larger than 1 but not very large. Under those conditions, compared to the dedicated system, pooling *some* of the dedicated queues improves social welfare. The reason is that with such a change, the system gets more congested, but the system's throughput considerably increases without a large increase in its average sojourn time. On the other hand, under the same conditions, if all of these partially-pooled queues are further pooled into a single queue, the system gets even more congested, leading to a considerable increase in the average sojourn time compared to the partially pooled case. In that case, the increase in throughput (due to complete pooling) is not large enough to dominate the effect of the aforementioned increase in the average sojourn time on social welfare. As a result, partial pooling maximizes social welfare when R/c is moderate, and ρ is strictly larger than 1 but not very large.

2.3.5 Unobservable System

Different from the setting in Section 3.1, the queue length information or real-time expected delay information is not available to customers in the unobservable system. Thus, this section is related to the literature that studies strategic, delay-sensitive customers that cannot observe the queue length but make decisions based on steady-state information. (Littlechild, 1974), (Edelson and Hilderbrand, 1975) and (Mendelson, 1985) are among the first who analyze single-server queueing models in unobservable settings. There is a rich literature that extends these papers in various dimensions. (See, e.g., Afèche (2013), Yang et al. (2017) and Ravner and Shamir (2017) for some of the novel problems studied in this context.) (Hassin and Haviv, 2003) include an excellent review of relevant papers that predate 2003. (Hassin, 2016a) includes other relevant papers that are published after 2003.

This section offers two key insights: If queue length information is not available to customers, (i) the pooled system always results in larger social welfare than the dedicated system when the service is free of charge, and (ii) the pooled system still dominates the dedicated system in terms of social welfare (revenue) when a fee is set to maximize social welfare (the provider's revenue).

We now explain our formulation. Arriving customers decide whether to join or balk based on the potential arrival rate Λ_j and average sojourn time in the system $j \in \{d, p\}$. A customer's joining/balking strategy is determined by a joining probability q_j for $j \in \{d, p\}$; a customer joins the queue with probability q_j and balks with probability $(1-q_j)$. The unique equilibrium in this setting is characterized and explained at the beginning of Appendix A.12.1 of the Electronic Companion. Let $\widehat{SW}_j$ and $\widehat{W}_j$ represent the equilibrium social welfare and average sojourn time in the system $j \in \{d, p\}$, respectively. Moreover, let $\widehat{L}_p$ the long-run average number of customers in the pooled system, and $\widehat{L}_d$ be the long-run average number of customers in one of the N dedicated sub-systems in equilibrium. Then,

$$\widehat{SW}_j = \begin{cases} \left(R - c\widehat{W}_j\right)\widehat{\lambda}_{e,j} = R\widehat{\lambda}_{e,j} - c\widehat{L}_j & \text{if } j = p \\ \left(R - c\widehat{W}_j\right)\widehat{\lambda}_{e,j}N = \left(R\widehat{\lambda}_{e,j} - c\widehat{L}_j\right)N & \text{if } j = d, \end{cases} \tag{2.19}$$

where $\widehat{\lambda}_{e,p}$ is the equilibrium effective arrival rate in the pooled system and $\widehat{\lambda}_{e,d}$ is the equilibrium effective arrival rate in one of the N dedicated sub-systems.[11]

[11] It is perhaps worth noting that (2.19) and (2.3) are different as (2.19) is concerned with the equilibrium performance measures, such as $\widehat{W}_j$, $\widehat{\lambda}_{e,j}$ and $\widehat{L}_j$.

Proposition 6. *In equilibrium, the unobservable pooled system results in larger social welfare (and smaller average sojourn time) than the unobservable dedicated system.*

The proof of this result is presented for a more general case with any given fixed fee $f \geq 0$. Thus, Proposition 6 is valid in a more general setting. Proposition 7 shows that the result extends to the case with heterogeneous service rewards.

Proposition 7. *Suppose that the customer service reward is distributed with a general distribution function $G(\cdot)$ defined on any bounded support. Then, in equilibrium, the unobservable pooled system results in larger social welfare (and smaller average sojourn time) than the unobservable dedicated system, i.e., $\widehat{SW}_p \geq \widehat{SW}_d$.*

Below, we will further check the robustness of the result in Proposition 6 under the following two formulations:

(i) *Welfare-maximizing fee*: For each system $j \in \{d, p\}$, a fee f_j is chosen to maximize the social welfare:

$$\max_{f_j \geq 0} \widehat{SW}_j \doteq (R - c\widehat{W}_j(f_j))N\lambda\hat{q}_j(f_j), \quad j \in \{d, p\}. \tag{2.20}$$

(ii) *Revenue-maximizing fee*: For each system $j \in \{d, p\}$, a fee f_j is chosen to maximize the service provider's revenue:

$$\max_{f_j \geq 0} \widehat{RV}_j \doteq N\lambda\hat{q}_j(f_j)f_j, \quad j \in \{d, p\}. \tag{2.21}$$

In both (2.20) and (2.21), $\hat{q}_j(f_j)$ represents the equilibrium joining probability for $j \in \{d, p\}$.

Proposition 8. **(a)** *Under formulation* (2.20)*, the unobservable pooled system outperforms the unobservable dedicated system in terms of the equilibrium social welfare at the welfare-maximizing fee, i.e., $\widehat{SW}_p^* \geq \widehat{SW}_d^*$.*

(b) *Under formulation* (2.21)*, the unobservable pooled system outperforms the unobservable dedicated system in terms of the equilibrium social welfare at the revenue-maximizing fee, i.e., $\widehat{SW}_p^{**} \geq \widehat{SW}_d^{**}$.*

It is straightforward to extend Proposition 8-(a) to the case with heterogeneous service rewards, and hence the proof of this extension is omitted. An immediate corollary of Proposition 8-(b) is that under

formulation (2.21), the maximum revenue under the unobservable pooled system is larger than that under the unobservable dedicated system in equilibrium. This is consistent with numerical observations by Ros and Tuffin (2004) that consider a queueing system with two "divisible" servers. These results and Propositions 6 through 8 suggest that in our formulation, the observability of queue (or customers having access to their real-time expected delay information) is a necessary condition for the dedicated system to outperform the pooled system in terms of social welfare.

The contrast between the results in the observable versus unobservable queue settings (i.e., Theorem 1 versus Propositions 6 through 8) can be explained as follows. In the observable system, each customer decides whether to join the system or not based on her *own* expected sojourn time. Specifically, each customer joins if and only if her own expected sojourn time is (weakly) smaller than the longest one, which is experienced by the customer who joins at the balking threshold. This joining behavior can lead to large congestion in the system. In contrast, when the system is unobservable, every customer makes her joining decision based on the equilibrium system state, i.e., average sojourn time across *all* customers that join the system. In equilibrium, every customer's joining probability is such that the effective arrival rate is always strictly smaller than the service rate. This is true even at large potential load. As a result, a key driver of Theorem 1, i.e., convex increasing stationary distribution of number of customers in the system, does not exist in the unobservable queue setting. In fact, because the effective system load is always strictly smaller than 1 and there is no queue capacity in the unobservable system, pooled and dedicated systems in the unobservable system behave similar to the ones studied by Smith and Whitt (1981). Hence, the classical benefit of pooling is recovered in the unobservable setting.

2.3.6 Observability as a System Feature

There could be practical scenarios in which running a system with an observable or unobservable queue are both feasible options. In such cases, a system can be run in one of the following four alternative ways: pooled observable, pooled unobservable, dedicated observable and dedicated unobservable. Considering these four alternatives, we proved Proposition 9 and Lemma 4. These results and the discussion following them complement the literature that studies an M/M/1 setting to understand if revealing queue length information improves the social welfare (see, for instance, (Hassin, 1986), Hassin and Roet-Green (2017b) and Hu et al. (2018)).

Proposition 9. **(a)** *When there is no service fee, the observable pooled system results in larger social welfare than both the unobservable pooled system and the unobservable dedicated system in equilibrium.*
(b) *When a fee is set to maximize the social welfare in each system, the observable pooled system results in the maximum social welfare among the four systems.*

Lemma 4. *When a fee is set to maximize the service provider's revenue in each system, the unobservable pooled system results in maximum social welfare among four systems, if R/c is moderate and ρ is not too large (i.e., $(N+1)/(N\mu) < R/c < \left((1+\rho)^2/\lambda + 1/\mu\right)$ and $\lambda < \bar{\lambda}$ for some constant $\bar{\lambda}$).*

A key implication of Proposition 9-(a) is that when there is no service fee, hiding queue length never improves the social welfare. Furthermore, Theorem 1-(a) and Proposition 9-(a) imply that when there is no service fee, the observable dedicated system results in maximum social welfare among the four systems if the conditions in (2.11) hold. By Proposition 9-(b), when all of the four systems are feasible options and a service fee is set to maximize the social welfare, hiding queue length cannot be welfare-maximizing. Lemma 4 shows that hiding queue length information can be welfare-maximizing when the service fee is set to maximize the provider's revenue. In the same setting, one can also show that among the four systems, the observable pooled system results in maximum social welfare when ρ and R/c are large.

2.4 Concluding Remarks

Our paper provides key insights for service management. Theorems 1 through 3 suggest that the pooling option should be evaluated very carefully in queueing systems as pooling queues may significantly decrease the social welfare and consumer surplus by considerably increasing the average sojourn time. Services with large benefit and large potential load are particularly prone to the potential harm of pooling when the queue length information is available to customers and pricing control is not a feasible option. For these type of services, the magnitude of the performance loss due to pooling can be even larger in large scale systems.

When pooling queues is inevitable and pricing control is not feasible for an observable queueing system, there could be other operational levers to improve social welfare and consumer surplus in the pooled system. In some practical settings, the number of servers could be a feasible operational lever. By changing the number of servers in the pooled system, one might improve the social welfare and the customer surplus under the pooled system. In fact, our numerical studies demonstrate that when the dedicated system outperforms the pooled system, by sufficiently increasing the number of servers in the pooled system, the performance

of the pooled systems achieves or exceeds the performance of the dedicated system. We also numerically observe that the minimum number of additional servers required in the pooled system to achieve or exceed the dedicated system's performance increases with the potential system load. This means that, when there is an increase in the potential system load, more server addition is necessary for the pooled system to perform as good as the dedicated system.

Another operational lever for performance improvement in the pooled system could be limiting the queue length under pooling. Our numerical studies suggest that when the potential system load is large, the social welfare and consumer surplus in the pooled system can be improved by choosing an appropriate buffer size for the pooled system. Such queue length control improves the system performance by mitigating over-utilization in the system, especially when the potential system load is very large. Our further numerical studies yield that the social welfare (and consumer surplus) in the pooled system is increasing and then decreasing with the buffer size. This means that the appropriate buffer size for the pooled system is a moderate one, which is typically smaller than the maximum number of customers in the dedicated system. Note that in the context of customer service, not being able to receive service from a particular channel, say, a call center, is not always equivalent to giving up service entirely. In various practical settings (e.g., e-commerce), a customer who does not join the call center queue can still receive service via a less desirable alternative channel such as a web form or e-mail. The strategy of trying to meet some of the customer service demand through these less desirable alternatives instead of adding servers to the call center could be a reason why some call centers might operate with a large potential load.

In certain practical settings, the observability of queue itself can be an operational lever. However, our analysis shows that when there is no service fee and pooling queues is inevitable, hiding the queue length or the real-time expected delay information never improves social welfare or consumer surplus.

This paper studies a setting where each customer gains the same reward R upon service completion. Our extensive numerical study shows that our main insights extend if customers' rewards are allowed to be different. As explained in Section 3.0.2, the literature suggests that the heterogeneity in jobs tends to benefit the dedicated system over the pooled system. Although these papers study settings that are very different than ours, our numerical observation that the dedicated system can continue to outperform pooled system when the customer reward is heterogeneous is consistent with the literature.

CHAPTER 3

Dynamic Learning and Rational Customers in Services

In this chapter, we consider the statistical learning of customers when they do not have full information of the service speed. They can learn the service speed based on their observations in the system and behave as forward-looking customers. We study the effect of learning and the operational implications for the service provider.

3.0.1 Summary of Main Results and Insights

We study three alternative policies for the strategic customers in a single-server queue: the naive policy, the myopic policy and the forward-looking policy.

Proposition 11 through 13 show that all the three alternative policies have threshold structures, and the forward-looking policy has smaller belief thresholds compared to the myopic policy and the naive policy. It implies that the forward-looking customer is more likely to join the queue when the true service rate is unknown to the customer. Proposition 14 characterizes the joining threshold of customers with different policies.

Proposition 18 shows that the expected value of learning is non-monotone in queue position assuming all the other customers are simple. The value is maximal at moderate queue position. It also shows that the expected value of learning is non-monotone in service reward and waiting cost per unit of time, and it is maximal at intermediate value. Proposition 19 compares the alternative policies given service rate is high or low.

We extend our model by allowing all the customers in the system belong to the same type. Proposition 15 through 17 compare the throughput of the system with each type of customers. We also consider the case where customers are fully rational. They can incorporate the estimates of the abandonment probability of all other customers ahead in next time period when they use dynamic programming to make decision.

We conduct extensive simulations to study the impact of fully rational behavior on system performance and study the value of learning in this complicated case.

3.0.2 Relevant Literature

Our work also contributes to the literature on strategic customers. (Hassin and Haviv, 2003) and (Hassin, 2016b) provide a comprehensive literature review in this area. The literature usually assumes all parameters are common knowledge, but it is not the case for the majority of the time. Customers can also learn over time. We review a few papers closely related to our work. (Cui and Veeraraghavan, 2016) consider an observable queue where customers do not know the true service rate and have different beliefs. It is shown that information revealing can decrease the social welfare. (Hassin and Roet-Green, 2017a) consider a model that customers have three options: join, balk, or inspect the queue length with a cost in an unobservable single-server queue. They characterize the optimal disclosure policy for revenue maximization and social welfare maximization. (Veeraraghavan et al., 2018) studied a model where customers learn the service rate distribution and estimate their expected remaining sojourn time based on their posterior belief of the service rate. Customers are not forward-looking and they make abandonment decisions based on the expected utility.

Exogenous abandonments have been studied in the literature. (Gans et al., 2003b) and (Ward, 2012) provide literature review on call centers with chapters on asymptotic analysis of queueing systems with customers' abandonment. (Whitt, 1999) studies the impact of information on state and remaining service time of all customers on a M/M/s/r queueing system, where the delay tolerance is exponentially distributed with a fixed rate. (Jouini et al., 2009) and (Jouini et al., 2011) consider call center models where patience time follows exponential distribution. (Garnett et al., 2002) and (Whitt, 2004) consider the heavy-traffic approximation for queues with abandonments. Endogenous abandonments are also studied in the literature. (Mandelbaum and Shimkin, 2000) considers heterogeneous customers whose patience time depends on the customers' belief about the distribution of the waiting time. (Afeche and Sarhangian, 2015) analyzes the equilibrium abandonment strategy of low-priority class customer for an observable queue with two-class priority customers. The impact of waiting time information is also studied in the literature. (Guo and Zipkin, 2007) and (Armony et al., 2009) consider the impact of delay announcement upon arrival.(Ata and Peng, 2017) consider endogenous abandonment of customers who are forward-looking. (Zohar et al., 2002) considers an invisible multi-server queue where customers' patience time distribution depends on the mean

waiting time in queue. (Hassin and Haviv, 1995) considers equilibrium strategies of impatient customers where the service reward drops to 0 if the waiting time exceeds a threshold. (Aksin et al., 2007) surveys the recent literature on the operation management of call centers.

Forward-looking behavior is also considered in the literature. (Yu et al., 2016) empirically study the impact of delay information on customers' strategic behaviors using the data from a call center. (Emadi and Swaminathan, 2017) empirically study the abandonment behavior using the data from a bank call center. They consider a model where customers update their beliefs on the waiting time distribution parameters through waiting experience by Bayes Rule, and apply the optimal stopping model to make abandonment decision. In our model, customers update their belief about service rate in a Bayesian framework based on service completion information, and they solve their optimal stopping problem based on their queue position and belief. We also compare this forward-looking policy with other alternative policies: myopic policy and naive policy from the perspectives of customers and the service provider.

It is generally accepted that strategic customers' behavior will hurt the firm, but there are also observations that it can be beneficial. Since the seminar work of (Coase, 1972), which pointed out that the market power of the durable good monopolist will be eliminated when consumers anticipate future price changes and can delay their purchase, there is growing literature on the adverse effects of customers' forward-looking behavior ((Stokey, 1979), (Besanko and Winston, 1990), (Su and Zhang, 2008), (Cachon and Swinney, 2009), (Aviv and Pazgal, 2008), (Parlaktürk, 2012),(Liu and Zhang, 2013)). In contrast, in some cases it is shown that the forward-looking behaviors of customers can be beneficial. (Su, 2007) finds that the strategic waiting of low-value customers can increase the willingness of high-value customers to pay. (Swinney, 2011) finds that the strategic customer behavior can be beneficial for the firm because of avoiding restocking costs. (Li et al., 2014) points out that strategic customer behavior may not hurt revenue because it also drives up demand when driving down price. (Lin et al., 2018)) finds that the strategic behavior is always beneficial for the manufacturer because of higher sales quantity. We compare the forward-looking policy with myopic policy and naive policy, and identify the conditions when the forward-looking behavior is beneficial or it is detrimental for either the customer or the service provider.

3.0.3 Outline of the Section

The section is organized as follows. Section 3.1 describes the model. Section 3.2 gives the preliminary results on three types of policies we consider. Section 3.3 states our main results about the structural proper-

ties of the policies. Section 3.5 compares the three alternative policies given the service rate is either high or low when there is a single strategic customer in the queue. Section 3.4 extends the model such that everyone in the system is strategic and uses the same policy. It presents results on the throughput comparison between different policies.

3.1 Model

Consider a first-come-first-served (FCFS) queuing system with one server in a discrete-time setting. Time is indexed by $t = 1, 2, \ldots$. It takes S periods to serve a customer and S has a geometric distribution with parameter p:

$$\mathbb{P}(S = k) = (1 - p)^{k-1}p, \quad k = 1, 2, 3, \ldots. \tag{3.1}$$

At the beginning of period 1, a new customer arrives to the system with probability $\gamma \in (0, 1)$. The arriving customer (he) does not know the parameter p but has a binary prior belief about it: he believes that $p = p_H$ with probability α_0 and $p = p_L$ with probability $1 - \alpha_0$ where p_L and p_H are known constants that satisfy $0 < p_L < p_H < 1$. Upon arrival, the customer observes the number of customers in the system, and decides whether to join the queue or leave the system based on the number of customers in the system and his prior belief α_0. Initially, there are N_0 customers in the system. If the customer chooses to leave the system, he never comes back. If the customer chooses to join the line, then at the end of each period $t = 1, 2, 3, \ldots$, the customer decides whether to continue waiting or abandon the system based on his posterior belief at the end of period t, that is α_t, and the number of customers in the system in front of him at the end of period t, that is N_t.

Define the stochastic process $X \doteq \{X_t : t = 1, 2, \ldots\}$ such that

$$X_t \doteq \begin{cases} 1 & \text{if there is a service completion in period t} \\ 0 & \text{if the service is not completed in period t} \end{cases} \tag{3.2}$$

In period $t = 1, 2, \ldots$, if the customer waits, then he observes X_t. Using the observation X_t and α_{t-1}, the customer forms its posterior belief at the end of period t as follows.

$$\alpha_t = \mathbb{P}(p = p_H | \mathcal{F}_t) \overset{(\star)}{=} \mathbb{P}(p = p_H | \alpha_{t-1}, X_t). \tag{3.3}$$

Here, $\mathcal{F}_t$ represents the entire information available to the customer up to the beginning of period t. Please note that ($\star$) indicates that the belief is a sufficient statistic, i.e., α_{t-1} includes all the necessary information until the end of period $t-1$. We will use y as a generic notation for the customer's belief.

The cost of waiting in line to the customer is $c > 0$ for one period. The customer receives reward R as soon as he is in service. When the customer abandons the system, he receives zero payoff.

At each decision epoch, $A_t \in \{0, 1\}$, where 0 stands for stay and 1 stands for leave. It is adapted to $\mathcal{F}_t$. The stopping time τ is defined as below:

$$\tau = \min\{t : \text{customer abandons the queue or receives service at the beginning of time period t}\}. \quad (3.4)$$

An admissible policy for a customer is a non-anticipating finite stopping time τ, at which customer abandons the queue or receives service. The expected total benefit is:

$$\mathbb{E}\left[\sum_{t=0}^{\tau} r(N_t, A_t)|N_0 = n, \alpha_0 = y\right] \quad (3.5)$$

where $r(N_t, A_t)$ is the reward in time period t given the action is A_t. It is given as below:

$$r(N_t, A_t) = \begin{cases} -c & \text{if } A_t = 0, \\ R & \text{if } A_t = 1 \text{ and } N_t = 0, \\ 0 & \text{if } A_t = 1 \text{ and } N_t > 0. \end{cases} \quad (3.6)$$

3.2 Preliminary Analysis

The customer uses Bayesian rule to update his belief at the end of each time period $t = 1, 2, \ldots$. We can calculate the posterior belief according to the following lemma.

Lemma 5. *A customer's posterior belief process $\{\alpha_t, t = 1, 2, \ldots\}$ satisfies*

$$\alpha_t = \mathbb{P}(p = p_H|\alpha_{t-1}, X_t) = \frac{p_H^{X_t}(1-p_H)^{1-X_t}\alpha_{t-1}}{p_H^{X_t}(1-p_H)^{1-X_t}\alpha_{t-1} + p_L^{X_t}(1-p_L)^{1-X_t}(1-\alpha_{t-1})}, \quad t = 1, 2, \ldots. \quad (3.7)$$

3.2.1 Forward-Looking Customer

Forward-looking customer uses dynamic programming to choose policy in order to achieve maximal expected total benefit:

$$V_F(n,y) \doteq \max \mathbb{E}\left[\sum_{t=0}^{\tau} r(N_t, A_t)|N_0 = n, \alpha_0 = y\right]. \tag{3.8}$$

$V_F(n,\alpha)$ satisfies the following dynamic programming equation:

$$\begin{aligned} V_F(n,y) = \max\Bigg\{0, -c + (yp_H + (1-y)p_L)V_F\left(n-1, \frac{p_H y}{p_H y + p_L(1-y)}\right) + \\ (y(1-p_H) + (1-y)(1-p_L))V_F\left(n, \frac{(1-p_H)y}{(1-p_H)y + (1-p_L)(1-y)}\right)\Bigg\} \end{aligned} \tag{3.9}$$

subject to the boundary condition:

$$V_F(0,y) = R, \quad y \in [0,1]. \tag{3.10}$$

3.2.2 Myopic Customer

Myopic customer also updates his belief according to (3.7). Instead of using dynamic programming, he makes decision in order to maximize his expected total benefit based on current belief y_t and queue position N_t at the end of each time period $t = 1, 2, \ldots$.

The customer will join the system at the beginning of $t = 1$ if and only if

$$R - \left(\frac{n_0}{p_H}\alpha_0 + \frac{n_0}{p_L}(1-\alpha_0)\right)c \geq 0. \tag{3.11}$$

He will continue waiting at the end of each time period $t = 1, 2, 3, \ldots$ if and only if

$$R - \left(\frac{n_t}{p_H}\alpha_t + \frac{n_t}{p_L}(1-\alpha_t)\right)c \geq 0. \tag{3.12}$$

We can also get the optimality equation for the expected total benefit of myopic learner as below:

$$V_M(n,y) = \begin{cases} 0, \quad \text{if } R < \left(y\frac{n}{p_H} + (1-y)\frac{n}{p_L}\right)c, \\ -c + (yp_H + (1-y)p_L)V_M\left(n-1, \frac{p_H y}{p_H y + p_L(1-y)}\right) \\ +(y(1-p_H) + (1-y)(1-p_L))V_M\left(n, \frac{(1-p_H)y}{(1-p_H)y + (1-p_L)(1-y)}\right), \quad \text{otherwise.} \end{cases} \tag{3.13}$$

The boundary condition is as below:

$$V_M(0,y) = R, \quad y \in [0,1]. \tag{3.14}$$

3.2.3 Naive Customer

A naive customer makes joining decision based on the expected total benefit $V_N(n,y)$.

$$V_N(n,y) = \max\left\{0, R - \left(\frac{n}{p_H}y + \frac{n}{p_L}(1-y)\right)c\right\}. \tag{3.15}$$

Similar to myopic learner, naive customer will join the system at the beginning of $t = 1$ if and only if

$$R - \left(\frac{n_0}{p_H}\alpha_0 + \frac{n_0}{p_L}(1-\alpha_0)\right)c \geq 0. \tag{3.16}$$

Different from myopic learner, naive customer does not update his belief in consecutive time periods. Thus, once a naive customer has chosen to join the system, he will never abandon the service and he will leave the system once the service is complete.

3.3 Analysis

Lemma 6. (a) *There exists a function V_F that satisfies the dynamic programming equation (3.9) subject to the boundary condition (3.10) and can be attained by an admissible policy π^*, i.e.,* $\mathbb{E}^{\pi^*}\left[\sum_{t=0}^{T} r(N_t, A_t)|N_0 = n, \alpha_0 = y\right] = V_F(n,y)$.

Lemma 6 indicates that there exists an optimal solution in our dynamic programming problem, and it can be attained by an optimal policy. Thus, forward- looking customers can determine their policy using the optimality equation in (3.9).

Proposition 10. *A forward-looking customer's optimal expected total benefit* $V_F(n, y)$ *is*

(a) *decreasing in* n *for any given* $y \in [0, 1]$, **(b)** *increasing in* y *for any given* $n \in \mathbb{N}$, **(c)** *increasing in* R, **(d)** *decreasing in* c. **(e)***convex in queue position* n *for given* y, $y \in [0, 1]$.

Proposition 10 describes the expected total benefit of a forward-looking customer in n and y.

It is shown that the $V_F(n, y)$ is decreasing more slowly as n increases. Eventually, when n is sufficiently large, the forward-looking customer will not join the system and will get a zero expected total benefit.

The expected total benefit of the naive customer and the myopic learner are also decreasing in n and increasing in y, increasing in R and decreasing in c. In addition, the expected total benefit of the naive customer is piecewise linear.

Figure 3.1 below displays an example of the value function $V_F(n, y)$. As shown in Proposition 10, $V_F(n, y)$ is decreasing in n and increasing in y.

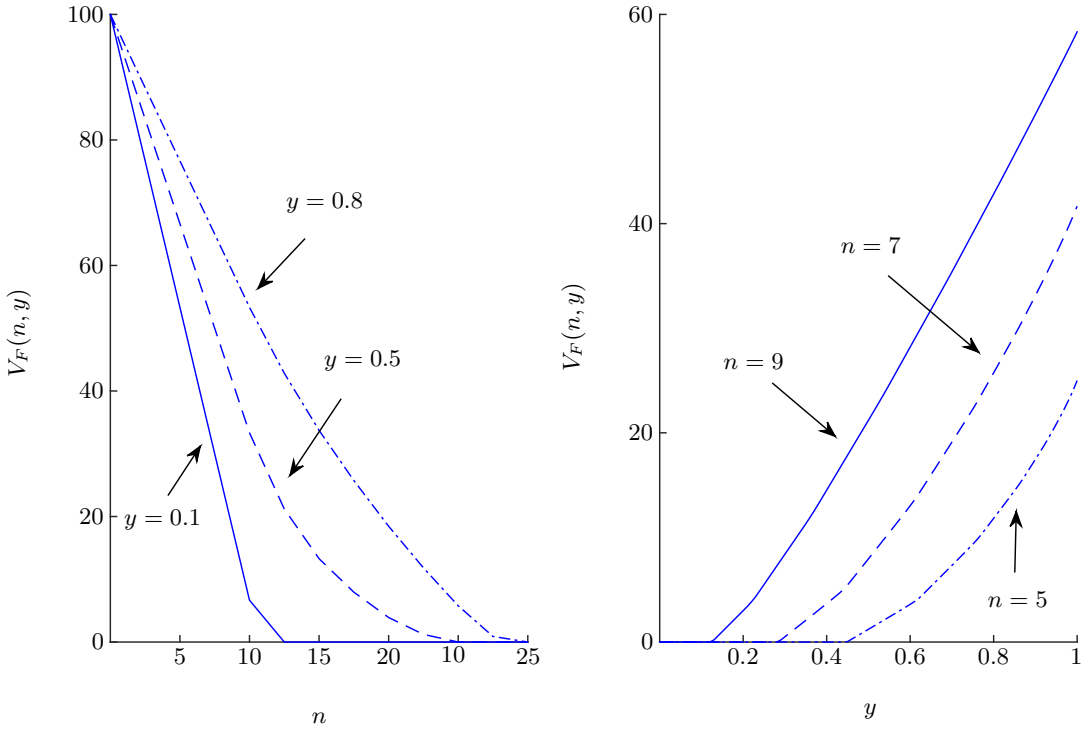

Figure 3.1: Value function when $R = 100$, $c = 5$, $p_H = 0.6$, $p_L = 0.2$.

Proposition 11 and 12 identify and compare the structure of the forward-looking policy, naive policy and myopic policy.

Proposition 11. *For any given n, there exist thresholds $\beta_F(n)$,$\beta_M(n)$,$\beta_N(n)$ such that*

(a) a forward-looking customer will wait if $y \geq \beta_F(n)$ and leave if $y < \beta_F(n)$.

(b) a myopic learner will wait if $y \geq \beta_M(n)$ and leave if $y < \beta_M(n)$. In addition, a naive customer will wait if $y \geq \beta_N(n)$ and leave if $y < \beta_N(n)$.

Proposition 12. *The belief threshold satisfy $\beta_M(n) = \beta_N(n) \geq \beta_F(n)$.*

Proposition 13. *For a forward-looking customer, the threshold $\beta_F(n)$ has the following properties:* **(a)** *$\beta_F(n) \in (0,1)$ if and only if $n \in (\bar{n}_1, \bar{n}_2)$, where*

$$\bar{n}_1 \doteq \frac{Rp_L}{c} \text{ and } \bar{n}_2 \doteq \frac{Rp_H}{c}. \tag{3.17}$$

(b) *$\beta_F(n)$ is increasing in n,* **(c)** *$\beta_F(n)$ is decreasing in R and increasing in c,* **(d)** *$\beta_M(n)$ or $\beta_N(n)$ satisfies that $\beta_M(n), \beta_N(n) \in (0,1)$ if and only if $n \in (\bar{n}_1, \bar{n}_2)$ and they are increasing in n, decreasing in R, and increasing in c.*

According to Proposition 10, $V_F(n,y)$ is decreasing in n, increasing in R and decreasing in c. Thus, we can get the results in parts (a) and (b) for optimal threshold. In part (c), It shows that the customer will not abandon the service for any $y \in [0,1]$ when n is sufficiently small, $i.e.$, $n < \frac{Rp_L}{c}$; $\beta_F(n) = 0$ when $n = \frac{Rp_L}{c}$; the customer will always abandon for any $y \in [0,1]$ when n is sufficiently large, $i.e.$, $n \geq \frac{Rp_H}{c}$.

Both the naive customer and the myopic learner use expected total benefit to make decision only based on current belief and queue position, thus they have the same threshold. The difference is that naive customer only makes joining decision based on the threshold, and he will never abandon the service in consecutive time periods. The forward-looking customer uses dynamic programming to make decision. Since the forward-looking policy can achieve larger expected total benefit for given n and y, thus his threshold is smaller.

Figure3.2 displays an example of the threshold comparison. $\beta_M(n) = \beta_N(n) = \beta_F(n) = 0$ when n is sufficiently small ($n \leq 4$). $\beta_F(n) \leq \beta_N(n) = \beta_M(n)$ for moderate n ($5 \leq n \leq 16$). All types of customers will not join when n is large ($n > 16$).

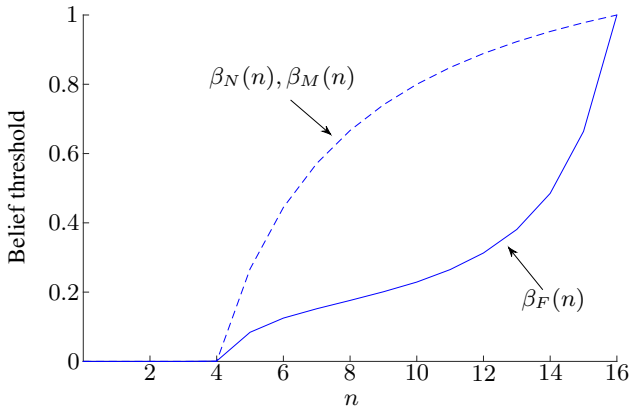

Figure 3.2: Belief threshold comparison when $R = 100$, $c = 5$, $p_H = 0.8$, $p_L = 0.2$.

Corollary 1. *When all other customers are naive customers, the forward-looking customer has a smaller abandonment probability than the myopic learner, i.e.,* $B_F(n) \leq B_M(n)$.

When all other customers behave like naive customers, we do not need to consider the possibility of abandonment of other customers. According to Proposition 12, $\beta_M(n) \geq \beta_F(n)$, for any $n \in \mathbb{N}$. Also we know that myopic learner and forward-looking customer have the same belief updating process. Consider any given sample path for myopic learner and forward-looking customer who join the system with n people ahead and initial belief y. If forward-looking customer abandons the service at the end of time period t, it implies that $y_t \leq \beta_F(N_t) \leq \beta_M(N_t)$, then myopic learner also abandons service if he has not abandoned earlier. Thus, the forward-looking customer has a smaller abandonment probability than the myopic learner.

Proposition 14. *For any given initial belief* y*, there exist* $n_S(y)$ *and* $n_F(y)$*,* $n_S(y) \leq n_F(y)$*, such that for* $n \leq n_S(y)$*, both forward-looking and naive customers will join; and for* n *in* $(n_S(y), n_F(y)]$*, forward-looking customer joins and naive customer does not join; for* $n > n_F(y)$*, neither type of customer will join.*

According to Proposition 12, $\beta_M(n) \geq \beta_F(n)$, for any $n \in \mathbb{N}$. Thus, the forward-looking customer is more likely to join.

3.4 Throughput comparison

We use the system throughput as the key performance metric, which is defined as the average number of service completions per time period for a given length of time. The formulas for the system with only

forward-looking customers and the system with only naive customers are as below:

$$\theta_F = \sum_{t=1}^{\tau} C_F^*(t)/\tau \tag{3.18}$$

and

$$\theta_N = \sum_{t=1}^{\tau} C_N^*(t)/\tau, \tag{3.19}$$

where $C_F^*(t)$ and $C_N^*(t)$ denote whether there is a service completion during time period t for the forward-looking customer and the naive customers, respectively.

Proposition 15. *When $\frac{R}{c} > \frac{1}{p_H - p_L}$, there exist $\underline{y}$ and $\bar{y}$ such that when initial belief $\alpha_0 \in (\underline{y}, \bar{y})$, the system with forward-looking customers results in strictly larger average throughput than the system with naive customers, i.e., $\theta_F(\tau) > \theta_N(\tau)$, for any finite time horizon τ when $\tau > 2\lfloor \bar{n}_1 \rfloor + 3$, where n_1 is defined in (3.17).*

Proposition 15 shows that when the service reward is not too small compared to waiting cost per unit time, and α_0 is moderate, then the system with forward-looking customers results in larger average throughput than the system with naive customers. This is due to the fact that forward-looking customers are more likely to join the system, and they rarely abandon under conditions of this proposition.

Proposition 16. **(a)** *There exists a $\bar{\bar{y}}$ such that when the initial belief $\alpha_0 \geq \bar{\bar{y}}$, the system with forward-looking customers results in strictly smaller average throughput than the system with naive customers, i.e., $\theta_F(\tau) < \theta_N(\tau)$, for any finite time horizon τ, where $\tau \geq 2\lfloor \bar{n}_2 \rfloor + t_1 + 1$ and $\bar{n}_2$ is defined as in (3.17).*

(b) *The system with forward-looking customers results same average throughput as the system with naive customers, i.e., $\theta_F = \theta_N$, if either*

$$(i)\ \frac{Rp_H}{c} - \frac{Rp_L}{c} < 1 \quad or \quad (ii)\ \alpha_0 < \underline{y} \tag{3.20}$$

where $\underline{y} = \beta_F(\lfloor \bar{n}_1 \rfloor + 1)$, $\bar{n}_1$ is defined in (3.17).

Proposition 16 shows that when initial belief α_0 is sufficiently large, the system with naive customer results in larger throughput than forward- looking customer. For sufficiently large initial belief, naive cus-

tomer and forward-looking customer have similar joining behavior, and naive customer will never abandon the service, which results in larger throughput.

It shows that the system with forward-looking customers has same throughput with the system with naive customers if (3.20). Under condition (i) in (3.20), both forward-looking customer and naive customer will always join/stay or always balk/abandon regardless of the belief for any queue position. Under condition (ii) in (3.20), initial belief is small enough that both forward-looking customer and naive customer will only join the system when $n \leq \bar{n}_1$, where the customer will never abandon regardless of the belief. Thus, they result in same throughput under these conditions.

Proposition 17. *The system with naive customers always results in larger average throughput than the system with myopic learners, i.e.,* $\theta_N \geq \theta_M$.

Proposition 17 shows that the system with naive customers always results in larger throughput than that with myopic learners. This is because naive customers and myopic learners have the same joining threshold, but myopic learners may abandon the service in consecutive time periods while naive customers never abandon.

We use simulation to compare the throughput of the system where all the customers are forward-looking with the system where all the customers are naive. We use 75 replications for 50000 time periods and the we choose 10000 warm-up periods. Table 3.1 below gives the throughput comparison for 9 pairs of relative reward R/c and initial belief α_0. Let $\gamma = \frac{\theta_F - \theta_N}{\theta_N} \cdot 100$. The tables below display γ. The percentage can be larger than 19% when $R = 30$ and $\alpha_0 = 0.35$.

Table 3.1: Throughput difference in percentage between the system with all the forward-looking customers and the system with all the naive customers

α_0	0.35	0.65	0.95
$R = 25$	[14.2085%, 14.2355%]	[4.6019%, 4.6151%]	[2.4257%, 2.4350%]
$R = 30$	[19.0781%, 19.1096%]	[5.9369%, 5.9541%]	[0.8989%, 0.9069%]
$R = 35$	[19.0676%, 19.0996%]	[5.9249%, 5.9423%]	[0.8998%, 0.9075%]

$c = 5, p = p_H = 0.6, p_L = 0.1, \lambda = 0.65$

Figure 3.3 and 3.4 show a example about the throughput difference in percentage of the system with a fraction w_1 of the customers are forward-looking with and the system where all the customers are naive. When the initial belief α_0 is moderate, the system with forward-looking customers results in larger throughput than the system with naive customers. The system with naive customers has larger throughput when the initial belief α_0 is large.

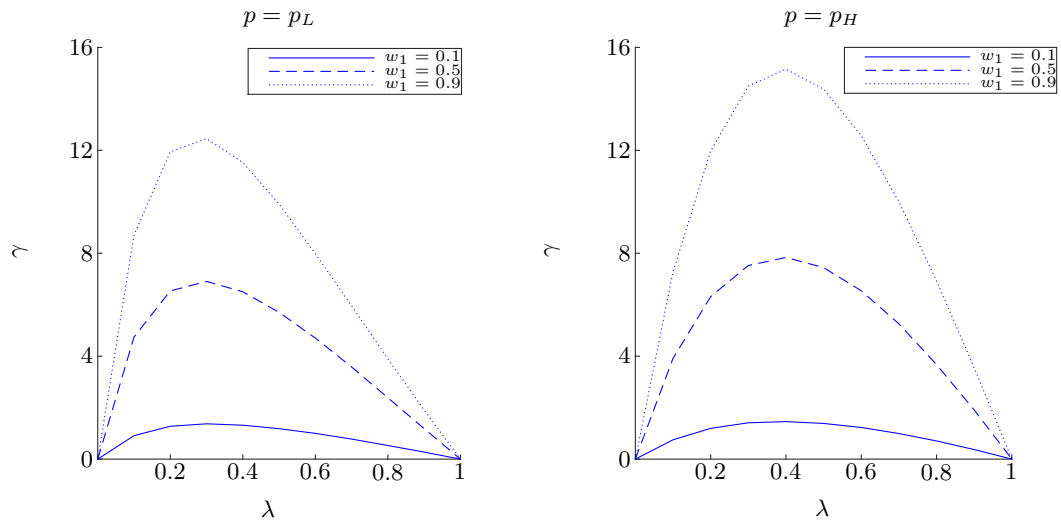

Figure 3.3: Throughput comparison when $R = 20$, $c = 5$, $p_H = 0.4$, $p_L = 0.2$, $\alpha_0 = 0.35$

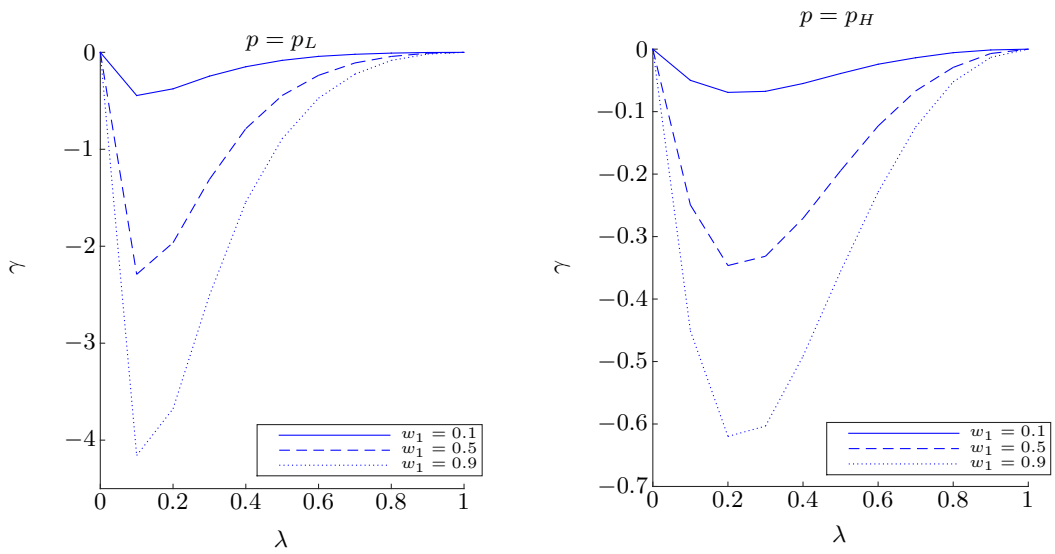

Figure 3.4: Throughput comparison when $R = 20$, $c = 5$, $p_H = 0.4$, $p_L = 0.2$, $\alpha_0 = 0.6$

Figure 3.5 below shows how the throughput difference in percentage change with the initial belief y. When the initial belief is small, the throughput of the system with forward-looking customers is larger than the system with naive customers. It means if the reputation of the service provider is bad, the forward-looking policy will benefit the service provider in terms of throughput. The reason is that the forward-looking customers are more likely to join the system than the naive customers

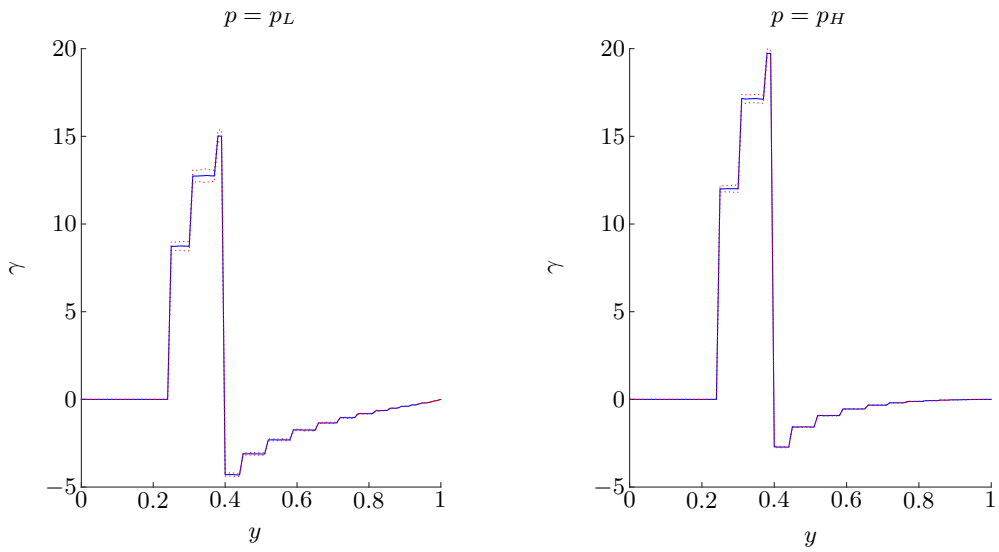

Figure 3.5: Throught comparison when $R = 20$, $c = 5$, $\lambda = 0.4$, $p_H = 0.4$, $p_L = 0.2$. The red dot lines show the 95% confidence interval

3.5 Value of Learning For Customers

Let $U_N^j(n, y)$, $j \in \{H, L\}$, denotes the expected total benefit of naive customer with belief y when there are n people ahead upon arrival given $p = p_j$. Since naive customer will never abandon the service once he has chosen to join the system, the expected total benefit is as below:

$$U_N^j(n, y) = \begin{cases} R - c\frac{n}{p_j} & \text{if } R - (y\frac{n}{p_H} + (1-y)\frac{n}{p_L})c > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{3.21}$$

where $j \in \{H, L\}$.

Let $U_F^j(n, y)$, $j \in \{H, L\}$, denotes the expected total benefit of the forward-looking customer with belief y when there are n people ahead of him upon arrival given $p = p_j$. Then we have the following

optimality equations.

$$U_F^j(n,y) = \begin{cases} 0 & \text{if } f(n,y) < 0, \\ -c + p_j U_F^j(n-1, g_1(y)) + (1-p_j) U_F^j(n, g_2(y)) & \text{if } f(n,y) \geq 0, \end{cases} \tag{3.22}$$

$$U_F^j(0,y) = R. \tag{3.23}$$

Here $f(n,y)$ is defined in (B.4). $g_1(y) = \frac{yp_H}{yp_H+(1-y)p_L}$ and $g_2(y) = \frac{y(1-p_H)}{y(1-p_H)+(1-y)(1-p_L)}$ as defined in (B.2) and (B.3).

The value of learning is defined as the gain in expected total benefit for the customer who is forward-looking instead of being naive given $p = p_H$ or $p = p_L$.

$$\Delta_j(n,y) = U_F^j(n,y) - U_N^j(n,y), \tag{3.24}$$

where $j \in \{H, L\}$.

The expected value of learning $\Delta(n,y)$ is as below:

$$\begin{aligned} \Delta(n,y) =& \Delta_H(n,y)y + \Delta_L(n,y)(1-y) \\ =& (U_F^H(n,y) - U_N^H(n,y))y + (U_F^L(n,y) - U_N^L(n,y))(1-y)) \\ =& U_F^H(n,y)y + U_F^L(n,y)(1-y) - (U_N^H(n,y)y + U_N^L(n,y)(1-y)) \\ =& V_F(n,y) - V_N(n,y) \end{aligned} \tag{3.25}$$

Proposition 18. *The expected value of learning for a forward-looking customer $\Delta(n,y)$ is* **(a)** *non-monotone with respect to queue length. In particular, there exists a threshold $\bar{n}$ such that the expected value of learning increases with the queue position n if $n < \bar{n}$; otherwise the expected value of learning decreases with the queue length,* **(b)** *increasing and then decreasing in R,* **(c)** *increasing and then decreasing in c.*

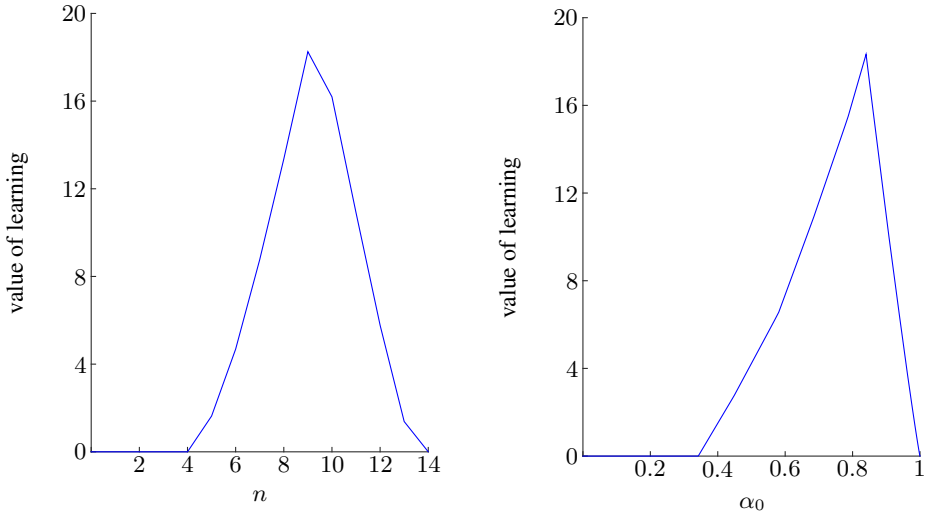

Figure 3.6: The expected value of learning when $R = 100$, $c = 5$, $p_H = 0.7$, $p_L = 0.2$. In the left panel, $\alpha_0 = 0.8$. In the right panel, $n = 10$

Proposition 18(a) is due to the fact that the expected benefit for forward-looking customer is convex in n, and the expected benefit of naive customer is piecewise linear in n. Also, we know that it is larger for forward-looking customer. Thus, the expected value of learning is non-monotone in n.

Proposition 18 shows that the value of learning is non-monotone in service reward R. It achieves maximum at some moderate R. The reason is that the expected total benefit of forward-looking customer is convex in R for any given n and y, while the value of naive customer is piecewise linear in R.

The proposition also states that the value of learning is non-monotone in waiting cost per unit time c. It achieves maximum value at some moderate c. The reason is that the expected total benefit of forward-looking customer is convex in c for any given n and y, while the value of naive customer is piecewise linear in c.

Is the ex-post value of learning still always positive? Next we study how the ex-post value of learning is affected by the initial belief and the queue position.

3.5.1 Impact of y

Proposition 19. *When there are n people ahead upon arrival and the initial belief $\alpha_0 = y$, if $\bar{y}_1(n) \in (0, 1)$, where $\bar{y}_1(n)$ is defined in (3.26), then there exists $\bar{y}_2(n) \in (0, 1)$ such that only the forward-looking customer*

joins when $\bar{y}_2(n) < y \leq \bar{y}_1(n)$,and all types of customers join when $y > \bar{y}_1(n)$,

$$\bar{y}_1(n) = \frac{1/p_L - R/cn}{1/p_L - 1/p_H}, \tag{3.26}$$

Furthermore,

(a) *if $p = p_H$ and $y > \bar{y}_1(n)$, the naive policy results in the maximum expected total benefit among all three policies. The value of learning is non-positive, i.e., $\Delta_H(n, y) \leq 0$. The loss from learning $-\Delta_H(n, y)$ is decreasing in y.*

(b) *if $p = p_L$ and $y > \bar{y}_1(n)$, the myopic policy results in the maximum expected total benefit among all three policies. The value of learning is non-negative, i.e., $\Delta_L(n, y) \geq 0$. The gain from learning $\Delta_L(n, y)$ is decreasing with y.*

(c) *if $p = p_H$ and $\bar{y}_2(n) < y \leq \bar{y}_1(n)$, the forward-looking policy results in the maximum expected total benefit among all three policies. The value of learning is non-negative, that is, $\Delta_H(n, y) \geq 0$. The gain $\Delta_H(n, y)$ is increasing with y.*

(d) *if $p = p_L$ and $\bar{y}_2(n) < y \leq \bar{y}_1(n)$, the naive policy and myopic policy result in the maximum expected total benefit among all three policies. The value of learning is non-positive, that is, $\Delta_L(n, y) \leq 0$. The loss $-\Delta_L(n, y)$ is increasing with y.*

Proposition 19 shows that being forward-looking is not better than being naive when the service rate is high given that both forward-looking customer and naive customer join the system. This is because naive customer does not update his belief and will never abandon the service once he has joined, which can be a good choice when service rate is high. The forward- looking customer will update his belief in each time period and his belief can decrease due to the failure of service completion. Since he only stays in the system for finite periods, his belief can be misleading due to the randomness of service completion. Thus, the forward-looking policy results in smaller expected total benefit. The result is completely opposite if the service rate is low.

Proposition 19 also shows that when belief is moderate, only forward-looking customer joins while naive customer does not join. It is better to use forward-looking policy when the service rate is high since the joining customer will receive positive expected total benefit. The result is opposite when service rate is low.

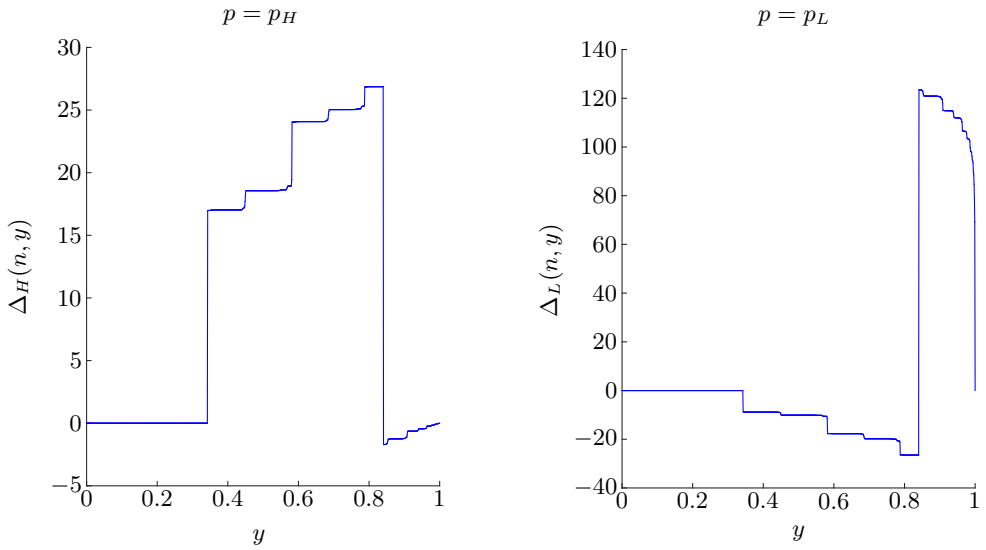

Figure 3.7: Value of Learning when $R = 100$, $c = 5$, $p_H = 0.7$, $p_L = 0.2$, $n = 10$.

Figure 3.7 pictures a numerical example for the value of learning as y increases for $n = 10$. When y is sufficiently small, both the forward-looking customer and the naive customer will not join the queue and thus the value of learning is 0. For moderate y, only the forward-looking customer joins. The forward-looking customer gains from learning and the gain is increasing in y when $p = p_H$. In contrast, the gain from learning is negative and the loss is increasing in y when $p = p_L$. When y is large, both the forward-looking customer and the naive customer join. The value of learning is negative and the loss is decreasing in y when $p = p_H$. It is positive and the gain is decreasing in y when $p = p_L$.

3.5.2 Impact of n

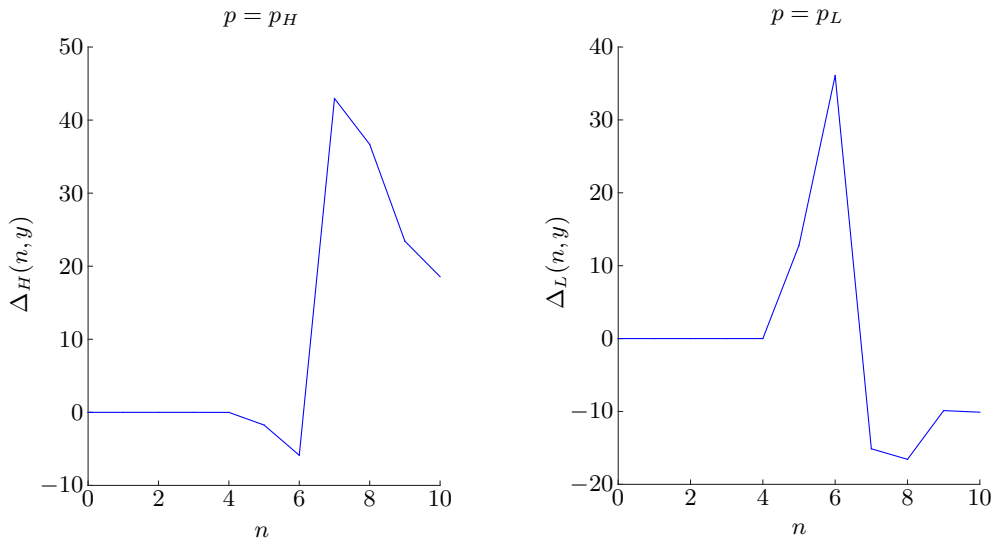

Figure 3.8: Value of Learning when $R = 100$, $c = 5$, $p_H = 0.7$, $p_L = 0.2$, $y = 0.5$.

Figure 3.8 pictures a numerical example for value of learning as n increases given $y = 0.5$. Specifically, the left panel pictures the value of learning when the true service rate is high. There is no value of learning when n is small ($n \leq 4$) since both types of customers will never abandon the service regardless of the belief. When n is moderate ($n = 5$ and $n = 6$), both types of customers join, the value of learning is negative since it is better to stay in the queue when the service rate is high in this range of n. When n is large ($n \geq 7$), only the forward-looking customer joins the queue while the naive customer balks. The value of learning is positive in this case, and it is decreasing with n. The right panel pictures the value of learning when the true service rate is low. In contrast, the value of learning is positive when n is moderate ($n = 5$ and $n = 6$) since it is better to leave the system in this range of n given $p = p_L$. The value of learning is negative when n is large ($n \geq 7$) since only the forward-looking customer joins and the expected total benefit is negative given $p = p_L$. The loss of learning is non-monotone in n in this example.

We also consider the value of learning when all the others are forward-looking by numerical studies. Even when all the others are forward-looking, the structure will be preserved. The value of learning is non-monotone and unimodal.

3.6 An Extension and Simulation Study

3.6.1 Fully Rational Customers

In previous sections, we considered customers that use Markov decision process to make joining/abandonment decisions assuming all the other customers do not abandon the service. In this section, we extend our model by considering the fully rational customers who can also incorporate others' abandon behaviors in their decision making process.

We consider a system with both fully rational customers and naive customers. The fraction of fully rational customers is w_1. The customers do not have full information about this fraction, and they think the fraction is w_2.

Recall that $\bar{n}_1 = \frac{Rp_L}{c}$. For customers in queue position n (n customers ahead in the system, including the one in service), $n \leq \bar{n}_1$, they will never abandon by the proof of Proposition 13. So we focus on the decision policy of customers in queue position $n \geq \bar{n}_1 + 1$.

Let $a_i(y)$ denote the optimal action of the customer in position i with belief y, where 0 stands for leave and 1 stands for stay.

For a customer in queue position n (n customers ahead), he assumes the customer in position i ($i = 0, 1, 2, \ldots, n-1$) has a posterior belief with uniform distribution over $\mathcal{B}_i$, where $\mathcal{B}_i \doteq \{y \in [0,1] : a_i(y) = 1\}$. Let $p(i)$ denotes the abandonment probability of the customer in position i if the service is not completed in the following time period given this customer is fully rational.

$$p(i) = \mathbb{P}\left(\frac{(1-p_H)y}{(1-p_H)y + (1-p_L)(1-y)} \notin \mathcal{B}_i | y \in \mathcal{B}_i\right) = \frac{\mathbb{P}\left(\frac{(1-p_H)y}{(1-p_H)y+(1-p_L)(1-y)} \notin \mathcal{B}_i, y \in \mathcal{B}_i\right)}{\mathbb{P}(y \in \mathcal{B}_i)}. \tag{3.27}$$

Let Z_i denotes the number of abandonments among the first i customers in the following time period if the service is not completed, and we have

$$\mathbb{P}(Z_i = 0) = 1, \quad \text{if } i \leq \bar{n}_1 + 1. \tag{3.28}$$

$$\mathbb{P}(Z_i = 0) = \mathbb{P}(Z_{i-1} = 0)(1 - p(i)w_2), \quad \text{if } i \geq \bar{n}_1 + 2 \tag{3.29}$$

$$\mathbb{P}(Z_i = j) = \mathbb{P}(Z_{i-1} = j)(1 - p(i)w_2) + \mathbb{P}(Z_{i-1} = j - 1)p(i)w_2, \quad \text{if } i \geq \bar{n}_1 + 2, 1 \leq j \leq i - (\bar{n}_1 + 1). \tag{3.30}$$

Denote the value function in this case as $V_R(n, y)$. The optimality equation is as below:

$$\begin{aligned} V_R(n,y) = \max \Bigg\{ & 0, -c + (yp_H + (1-y)p_L) V_R\left(n-1, \frac{p_H y}{p_H y + p_L(1-y)}\right) \\ & + (y(1-p_H) + (1-y)(1-p_L)) \Bigg(P(Z_n = 0) V_R\left(n, \frac{(1-p_H)y}{(1-p_H)y + (1-p_L)(1-y)}\right) \\ & + P(Z_n = 1) V_R\left(n-1, \frac{(1-p_H)y}{(1-p_H)y + (1-p_L)(1-y)}\right) \\ & + P(Z_n = 2) V_R\left(n-2, \frac{(1-p_H)y}{(1-p_H)y + (1-p_L)(1-y)}\right) \\ & + \ldots\ldots \\ & + P(Z_n = n - \bar{n}_1 - 1) V_R\left(\bar{n}_1 + 1, \frac{(1-p_H)y}{(1-p_H)y + (1-p_L)(1-y)}\right)\Bigg)\Bigg\}. \end{aligned} \tag{3.31}$$

$$V_R(0, y) = R \qquad \text{for } y \in [0, 1]. \tag{3.32}$$

3.6.2 Numerical Study for the Optimal Policy

We numerically calculate the value function and thresholds sequentially for the fully rational policy by the value iteration algorithm using equation (3.27) through (3.31). We find the policy of the fully rational customers is of threshold type. Denote the belief threshold for the customer in the nth position as $\tau_R(n)$. It has similar properties with the belief threshold of the forward-looking customers.

Figure 3.9 pictures a numerical example on how the thresholds $\tau_{ss}(n)$ change as w_2 changes. It indicates that the threshold is smaller with larger w_2. When w_2 is larger, the fully rational customer expect that the number of abandonments ahead of her is larger in the following period and she has a larger expected total benefit given current belief and queue position. Thus, the threshold is smaller with larger w_2.

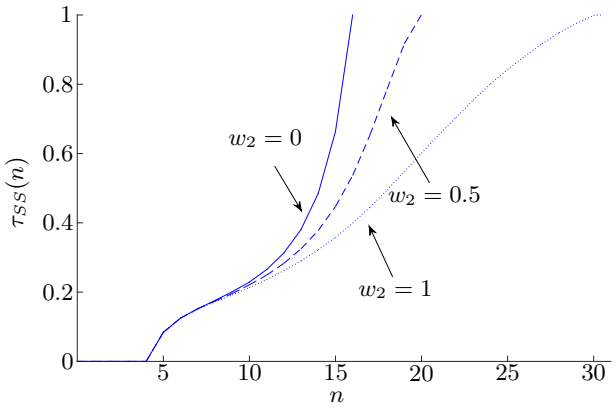

Figure 3.9: Threshold comparison when $R = 100$, $c = 5$, $p_H = 0.8$, $p_L = 0.2$.

Note that $w_2 = 0$ corresponds to the case that the customer assumes all the other customers are naive and the policy is equivalent to the forward-looking policy.

3.6.3 Simulation

We conduct simulation studies to evaluate the performance of the system with fully rational customers comparing to the system with only naive customers. We choose the long-run average throughput as the performance metric.

3.6.3.1 Throughput

Let θ_R denote the long-run average throughput of the system with both fully rational customers and naive customers, and θ_N denote that of the system with only naive customers. We use β_θ to denote the throughput difference in percentage and it is given as below:

$$\beta_\theta \doteq \frac{\theta_R - \theta_N}{\theta_N} 100. \tag{3.33}$$

Figure 3.10 displays $\beta_\theta(\lambda)$ with λ as an argument for three choices of w_1 and the initial belief is moderate. The difference in percentage is positive for either $p = p_H$ or $p = p_L$, and usually achieves maximal at a moderate λ, and it increases as w_1 increases.

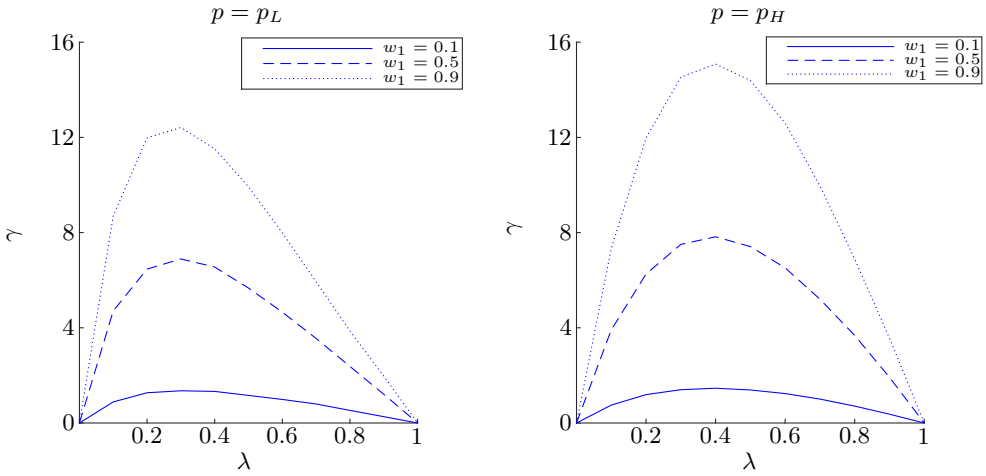

Figure 3.10: Throughput comparison when $R = 20$, $c = 5$, $p_H = 0.4$, $p_L = 0.2$, $w_2 = w_1$, $\alpha_0 = 0.35$.

3.6.4 Value of Learning

We compare the system with naive customers and the system with a combination of naive customers and fully rational customers. We define the value of learning $\widehat{\Delta}_j(n, y), j \in \{L, H\}$, as the difference of expected utility of customers with given position n and initial belief y for these two systems. In the system with a combination of fully rational customers and naive customers, let $\widehat{U}_R^j(n, y)$ and $\widehat{U}_S^j(n, y)$ denote the expected utility of a customer with queue position n and initial belief y for these two systems given $p = p_j$, respectively. The the value of learning is as below:

$$\widehat{\Delta}_j(n, y) = \widehat{U}_R^j(n, y) - \widehat{U}_S^j(n, y),\ j \in \{L, H\}. \tag{3.34}$$

It has similar patterns with the value of learning when customers are forward-looking instead of fully rational. Figure 3.11 below shows that the value of learning is monotone with the fraction of the fully-rational customers.

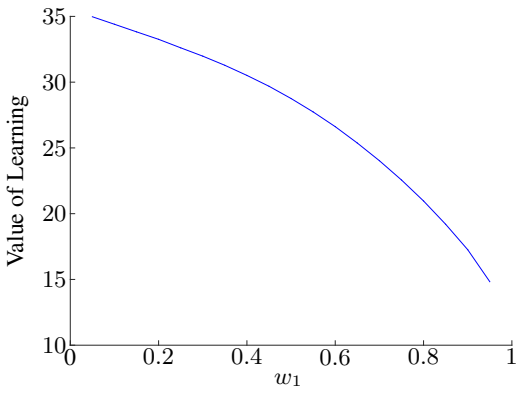

Figure 3.11: Expected Value of Learning when $R = 100$, $c = 5$, $p_H = 0.8$, $p_L = 0.1$, $\alpha_0 = 0.8$, $n = 6$, $\lambda = 0.9$, $w_2 = w_1$.

3.7 Concluding Remarks

This paper studies the dynamic learning behavior of customers when they do not have full information about the service rate. Customers can learn the service rate when they are waiting in the queue and use Markov decision policy to make abandonment decisions. We compare this policy with other heuristic policies and look at the value of such learning behaviors when there is only one strategic customer. The expected value of learning is positive and it achieves maximum at moderate queue position. In contrast, the ex-post value of learning can be negative given the service rate. We identify the conditions under which such learning behavior can hurt the customer's benefit and study how it is affected by the initial belief and the queue position.

We also compare the long-run average throughput for the system with different types of customers. It is commonly known that the customers being too smart are not good for the service provider. Surprisingly we find that the forward-looking customers can improve the system throughput. The improvement in percentage can be larger than 15%.

In addition, we consider the fully rational customers who can also consider the others' abandonment behavior when they make decisions. The optimal policy again has a threshold structure like the forward-looking policy. Our main results for the throughput comparison still carry through when customers are fully-rational according to the simulation.

APPENDIX A

PROOF OF RESULTS IN CHAPTER 2

A.1 Proofs of Lemmas 1 and 2 and a Supplementary Result

Proof of Lemma 1: Consider the pooled system described in Section 3.1. If there are already $n \geq N$ customers in the system, the expected time that an arriving customer spends in the system is $\bar{W}_p(n+1) = (n+1)/(N\mu)$. The reason is as follows. The arriving customer enters the service after $(n-N+1)$ customers ahead of her are served. Then, because there are N servers, it follows from standard probability arguments that the total expected time to complete the service of $(n-N+1)$ customers ahead of the arriving customer is $(n-N+1)/(N\mu)$. This gives the expected waiting time of the arriving customer in the queue. In addition to this, the expected service time of the aforementioned customer is $1/\mu$. Combining the expected waiting time in the queue and the expected service time, the expected time that the arriving customer spends in the system is $(n-N+1)/(N\mu) + 1/\mu = (n+1)/(N\mu)$. This and (2.2) imply that such an arriving customer joins the system if and only if $R - ((n+1)c/(N\mu)) \geq 0$ which is equivalent to $n \leq (RN\mu)/c - 1$. This implies that the maximum number customers in the system is equal to K where K is as defined in (2.6). To find the average sojourn time, we first identify the steady-state probability distribution of the number of customers in the system. Let π_i be the steady-state probability that there are i customers in the system. Because $K \geq N$ by (2.5), the balance equations of this system are the following: $N\lambda\pi_i = (i+1)\mu\pi_{i+1}$ for $i = 0, \ldots, N-1$ and $\pi_i\lambda = \pi_{i+1}\mu$ for $i = N, \ldots, K-1$. From these and the fact that $\sum_{i=0}^{K} \pi_i = 1$, we have

$$\pi_0 = \left(\sum_{i=0}^{N-1} \frac{N^i}{i!}\rho^i + \frac{N^N}{N!} \sum_{i=N}^{K} \rho^i \right)^{-1}, \tag{A.1}$$

$$\pi_i = \pi_0 N^i \rho^i / i! \quad \text{for } i = 1, \ldots, N \quad \text{and} \quad \pi_i = \pi_0 N^N \rho^i / N! \quad \text{for } i = N+1, \ldots, K. \tag{A.2}$$

Using these, we get the following expressions for the long-run average number of customers in the system and the throughput, respectively:

$$L_p = \sum_{i=0}^{K} \pi_i i = \frac{\sum_{i=0}^{N-1} \frac{N^i}{i!} i\rho^i + \frac{N^N}{N!} \sum_{i=N}^{K} i\rho^i}{\sum_{i=0}^{N-1} \frac{N^i}{i!} \rho^i + \frac{N^N}{N!} \sum_{i=N}^{K} \rho^i} \tag{A.3}$$

$$\lambda_{e,p} = (1 - \pi_K) N\lambda = \left(1 - \frac{\frac{N^N}{N!}\rho^K}{\sum_{i=0}^{N-1} \frac{N^i}{i!}\rho^i + \frac{N^N}{N!} \sum_{i=N}^{K} \rho^i}\right) N\lambda. \tag{A.4}$$

Because $W_p = L_p/\lambda_{e,p}$ by Little's law, (2.7) immediately follows from (A.3) and (A.4). Replacing (A.3) and (A.4) in place of L_p and $\lambda_{e,p}$, respectively, in the SW_p formula (2.3), we get (2.8). □

Proof of Lemma 2: Recall that the dedicated system consists of N separate sub-systems each with a line dedicated to one server. Suppose that a customer arrives to one of these dedicated queues and observes that there are n customers in that sub-system. Then, the expected time the arriving customer spends in that sub-system is $\bar{W}_d(n+1) = (n+1)/\mu$. This means by (2.2) that a customer joins the dedicated queue if and only if $R - (n+1)c/\mu \geq 0$ which is equivalent to $n \leq (R\mu)/c - 1$. Note that the later inequality implies that the maximum number of customers in each separate sub-system is k where k is as defined in (2.5). As a result, each separate sub-system can be considered as an $M/M/1/k$ system. Then, the long-run average number of customers and the throughput in one of the dedicated sub-systems are as follows, respectively (see Table 4 on page 149 of Sztrik (2012)).

$$L_d = \begin{cases} \frac{\rho[1-(k+1)\rho^k + k\rho^{k+1}]}{(1-\rho)(1-\rho^{k+1})} = \frac{\rho}{1-\rho} - \frac{(k+1)\rho^{k+1}}{1-\rho^{k+1}} & \text{if } \rho \neq 1 \\ \frac{k}{2} & \text{if } \rho = 1. \end{cases} \tag{A.5}$$

$$\lambda_{e,d} = \begin{cases} \lambda\left(1 - \frac{(1-\rho)\rho^k}{1-\rho^{k+1}}\right) = \lambda\left(\frac{1-\rho^k}{1-\rho^{k+1}}\right) & \text{if } \rho \neq 1 \\ \lambda\left(\frac{k}{k+1}\right) & \text{if } \rho = 1. \end{cases} \tag{A.6}$$

By Little's Law, $W_d = L_d/\lambda_{e,d}$. From this, (A.5) and (A.6), we get (A.7) below. Similarly, by substituting L_d and $\lambda_{e,d}$ respectively with (A.5) and (A.6) in the SW_d formula (2.3), we get (A.8) below.

$$W_d = \begin{cases} \frac{\rho-(k+1)\rho^{k+1}+k\rho^{k+2}}{\lambda(1-\rho)(1-\rho^k)} & \text{if } \rho \neq 1 \\ \frac{k+1}{2\lambda} & \text{if } \rho = 1, \end{cases} \tag{A.7}$$

$$SW_d = \begin{cases} \left(\frac{1-\rho^k}{1-\rho^{k+1}}\right) RN\lambda - \left(\frac{\rho}{1-\rho} - \frac{(k+1)\rho^{k+1}}{1-\rho^{k+1}}\right) Nc & \text{if } \rho \neq 1 \\ \frac{k}{k+1} RN\lambda - \frac{k}{2} Nc & \text{if } \rho = 1. \end{cases} \tag{A.8}$$

The fact that (A.7) and (A.8) are equivalent to W_d and SW_d expressions in the lemma, respectively, completes the proof. □

We now state and prove a supplementary lemma which we will use in the remainder of the Appendix.

Lemma 7. *Consider an $M/M/1/K$ queueing system (indexed by $j = s$) with the potential arrival rate λN and the service rate μN and K is as defined in (2.6). Suppose that there is no service fee as in Section 3.1. Then, (a) the throughput is*

$$\lambda_{e,s} = \begin{cases} N\lambda\left(\frac{1-\rho^K}{1-\rho^{K+1}}\right) & \text{if } \rho \neq 1 \\ N\lambda\frac{K}{K+1} & \text{if } \rho = 1. \end{cases} \tag{A.9}$$

(b) *The long-run average number of customers in the system is*

$$L_s = \begin{cases} \frac{\rho}{1-\rho} - \frac{(K+1)\rho^{K+1}}{1-\rho^{K+1}} & \text{if } \rho \neq 1 \\ \frac{K}{2} & \text{if } \rho = 1. \end{cases} \tag{A.10}$$

(c) *The average sojourn time is*

$$W_s = \begin{cases} \frac{\rho-(K+1)\rho^{K+1}+K\rho^{K+2}}{(1-\rho)(1-\rho^K)N\lambda} & \text{if } \rho \neq 1 \\ \frac{K+1}{2N\lambda} & \text{if } \rho = 1. \end{cases} \tag{A.11}$$

(d) *The social welfare is*

$$SW_s = \begin{cases} \left(\frac{1-\rho^K}{1-\rho^{K+1}}\right) RN\lambda - \left(\frac{\rho}{1-\rho} - \frac{(K+1)\rho^{K+1}}{1-\rho^{K+1}}\right) c & \text{if } \rho \neq 1 \\ \frac{K}{K+1} RN\lambda - \frac{K}{2} c & \text{if } \rho = 1. \end{cases} \tag{A.12}$$

Proof of Lemma 7: Because the ratio of the potential arrival rate to the service rate in the described system is $\frac{N\lambda}{N\mu} = \rho$, which is same as in each dedicated sub-system (which consists of one dedicated queue and one server), the balance equations for the described $M/M/1/K$ system are the same as the ones for the $M/M/1/k$ system analyzed in the proof of Lemma 2, with the exception that k must be replaced with K. Replacing k with K in $\lambda_{e,d}/\lambda$, we get $(1 - \tilde{\pi}_K)$ where $\tilde{\pi}_K$ is the steady-state probability of having K customers in the described $M/M/1/K$ system. From this and the fact that the throughput in $M/M/1/K$ system is equal to $(1 - \tilde{\pi}_K)N\lambda$, we complete the proof of part (a). Similarly, replacing k with K in L_d we get L_s in part (b). Because the average sojourn time is $W_s = \frac{L_s}{\lambda_{e,s}}$ by Little's Law, part (c) immediately follows. Finally, by replacing λ with $N\lambda$, μ with $N\mu$ and k with K in SW_d/N (from (A.8)), we get part (d). □

A.2 Proof of Theorem 1

A.2.1 Proof of Theorem 1 - Part (a):

Proposition 3 will show that if (2.13), then $W_d < W_s$ and $SW_d > SW_s$. This and the facts that $W_s < W_p$ and $SW_s > SW_p$ by Proposition 2 complete the proof of the claim in part (a). We provide the proofs of Propositions 2 and 3 in Appendices A.4 and A.5, respectively. □

A.2.2 Proof of Theorem 1 - Part (b):

We will first state and prove a lemma. Using this lemma, we will then prove parts (i) and (ii) under the condition ($\star$) in (2.12), and then we will prove same claims under the condition ($\star\star$) in (2.12).

Lemma 8. *Compared to the dedicated system, the pooled system results in strictly larger throughput than the pooled system, i.e.,* $\theta_d = N\lambda_{e,d} < \theta_p = \lambda_{e,p}$.

Proof of Lemma 8: Let b_d and b_p denote the balking probabilities in the dedicated and pooled systems, respectively. The proofs of Lemmas 1 and 2 imply that

$$b_d = \bar{\pi}_k = \frac{\rho^k}{1+\rho+\cdots+\rho^k} \quad \text{and} \quad b_p = \pi_K = \frac{\frac{N^N}{N!}\rho^K}{\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i + \frac{N^N}{N!}\sum_{i=N}^{K}\rho^i}, \tag{A.13}$$

where $\bar{\pi}_k$ is the stationary probability that there are k customers in one of the N sub-systems in the dedicated one. Based on these, observe that regardless of the value of ρ, we have

$$\begin{aligned}
&b_d - b_p \\
=&\frac{\rho^k}{\sum_{i=0}^{k}\rho^i} - \frac{\frac{N^N}{N!}\rho^K}{\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i + \frac{N^N}{N!}\sum_{i=N}^{K}\rho^i} \\
=&\frac{1}{(\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i + \frac{N^N}{N!}\sum_{i=N}^{K}\rho^i)\sum_{i=0}^{k}\rho^i}\left(\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^{i+k} + \frac{N^N}{N!}\sum_{i=N+k}^{K+k}\rho^i - \frac{N^N}{N!}\sum_{i=K}^{K+k}\rho^i\right) \\
=&\frac{1}{(\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i + \frac{N^N}{N!}\sum_{i=N}^{K}\rho^i)\sum_{i=0}^{k}\rho^i}\left(\sum_{i=0}^{N-2}\frac{N^i}{i!}\rho^{i+k} + \frac{N^N}{N!}\sum_{i=N+k-1}^{K+k}\rho^i - \frac{N^N}{N!}\sum_{i=K}^{K+k}\rho^i\right) &\text{(A.14)} \\
\geq&\frac{1}{(\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i + \frac{N^N}{N!}\sum_{i=N}^{K}\rho^i)\sum_{i=0}^{k}\rho^i}\left(\sum_{i=0}^{N-2}\frac{N^i}{i!}\rho^{i+k}\right) &\text{(A.15)} \\
>&0 &\text{(A.16)}
\end{aligned}$$

The equation (A.14) is because $\frac{N^{N-1}}{(N-1)!}\rho^{N+k-1} = \frac{N^N}{N!}\rho^{N+k-1}$. The inequality (A.15) follows from the fact that $N+k-1 \leq K$ since $N+k-1-K \leq N+k-1-Nk = (N-1)(1-k) \leq 0$. Because $\lambda_{e,p} = (1-b_p)N\lambda$ and $\lambda_{e,d} = (1-b_d)\lambda$, and the balking probability in the dedicated system is strictly larger than the one in the pooled system by (A.16), we have $\lambda_{e,d}N < \lambda_{e,p}$. □

Proof of Theorem 1 - Part (b) under the condition ($\star$) in (2.12): If $\frac{R}{c} < \frac{N+1}{N\mu}$, then $\frac{R\mu}{c} < \frac{N+1}{N}$ and $\frac{RN\mu}{c} < N+1$. This and (2.5) imply that $k=1$ and $K=N$. Thus, under the condition ($\star$), there is no waiting line and a joining customer immediately gets the service. As a result,

$$W_d = W_p = \frac{1}{\mu},$$

which is the claim in part (i). Recall the definition of the social welfare from (2.3):

$$SW_d = \lambda_{e,d} N(R - cW_d) = \lambda_{e,d} N \left(R - \frac{c}{\mu} \right) \quad \text{and} \quad SW_p = \lambda_{e,p}(R - cW_p) = \lambda_{e,p} \left(R - \frac{c}{\mu} \right).$$

Based on this, because $\lambda_{e,d} N < \lambda_{e,p}$ by Lemma 8, $SW_d < SW_p$. $\square$

Proof of Theorem 1 - Part (b) under the condition ($\star\star$) in (2.12): Define

$$\bar{\eta} \doteq (z_3 + 1)/\mu, \tag{A.17}$$

where

$$z_3 \doteq \inf \left\{ z \in \mathbb{R} : z > -(\ln(\rho)^{-1} + 3) \text{ and } (z+3)\rho^{z-1} < (N-1)/N \right\}. \tag{A.18}$$

In light of this, the outline of the remainder of the proof is as follows. First, we will state and prove Lemma 9 that shows the existence of the constant z_3 which will be used in the remainder of the proof. Then, we will show in Lemma 10 that if $\rho < 1$ and $k > z_3$, we have $L_p - NL_d < 0$. Finally, we will use this inequality to prove the claims in parts (i) and (ii) for $\rho < 1$ and $k > z_3$. This and the fact that $\frac{R}{c} > \frac{z_3+1}{\mu}$ implies $k > z_3$ complete the proof of part (b).

Lemma 9. *For $\rho < 1$, the constant z_3 defined in $(A.18)$ exists and it is finite.*

Proof of Lemma 9: Define $g_4(z) \doteq (z+3)\rho^{z-1}$. Then, note that the definition in (A.18) is equivalent to $z_3 \doteq \inf\{z \in \mathbb{R} : g_4(z) < (N-1)/N \text{ and } z > -1/\ln(\rho) - 3\}$. Observe that $g_4(\cdot)$ is strictly decreasing for $\rho < 1$ and $z > -\frac{1}{\ln(\rho)} - 3$ because

$$g_4'(z) = \rho^{z-1} + (z+3)\rho^{z-1}\ln(\rho) = \rho^{z-1}(1 + (z+3)\ln(\rho)) < 0. \tag{A.19}$$

In addition, by an application of L'Hopital's Rule, we have

$$\lim_{z \to \infty} g_4(z) = \lim_{z \to \infty} (z+3)\rho^{z-1} = 0. \tag{A.20}$$

From (A.19) and (A.20), the claim follows. $\square$

Lemma 10. *For $\rho < 1$ and $k > z_3$, the long-run average number of customers in the dedicated system, that is, NL_d, and the long-run average number of customers in the pooled system satisfy the following inequality:*

$$L_p - NL_d < 0. \tag{A.21}$$

Proof of Lemma 10: Recall L_p and L_d from (A.3) and (A.5). Then, we have

$$\begin{aligned}
&L_p - NL_d \\
&= \frac{\sum_{i=0}^{N-1} \frac{N^i}{i!} i\rho^i + \frac{N^N}{N!} \sum_{i=N}^{K} i\rho^i}{\sum_{i=0}^{N-1} \frac{N^i}{i!} \rho^i + \frac{N^N}{N!} \sum_{i=N}^{K} \rho^i} - N\frac{\rho\left(1-(k+1)\rho^k + k\rho^{k+1}\right)}{(1-\rho)(1-\rho^{k+1})} \\
&= \frac{1}{\left(\sum_{i=0}^{N-1} \frac{N^i}{i!} \rho^i + \frac{N^N}{N!} \sum_{i=N}^{K} \rho^i\right)(1-\rho)(1-\rho^{k+1})} \\
&\cdot \left[\left(\sum_{i=0}^{N-1} \frac{N^i}{i!} i\rho^i + \frac{N^N}{N!} \sum_{i=N}^{K} i\rho^i\right)(1-\rho)(1-\rho^{k+1})\right. \\
&\left. - N\rho\left(1-(k+1)\rho^k + k\rho^{k+1}\right)\left(\sum_{i=0}^{N-1} \frac{N^i}{i!} \rho^i + \frac{N^N}{N!} \sum_{i=N}^{K} \rho^i\right)\right]
\end{aligned} \tag{A.22}$$

Note that we have

$$\sum_{i=0}^{N-1} \frac{N^i}{i!} i\rho^i = \sum_{i=1}^{N-1} \frac{N^i}{i!} i\rho^i = \sum_{i=1}^{N-1} \frac{N^i}{(i-1)!} \rho^i = N\rho \sum_{i=1}^{N-1} \frac{N^{i-1}}{(i-1)!} \rho^{i-1} = N\rho \sum_{i=0}^{N-2} \frac{N^i}{i!} \rho^i,$$

$$\sum_{i=0}^{N-1} \frac{N^i}{i!} \rho^i + \frac{N^N}{N!} \sum_{i=N}^{K} \rho^i = \sum_{i=0}^{N-2} \frac{N^i}{i!} \rho^i + \frac{N^N}{N!} \sum_{i=N-1}^{K} \rho^i = \sum_{i=0}^{N-2} \frac{N^i}{i!} \rho^i + \frac{N^N}{N!} \frac{\rho^{N-1} - \rho^{K+1}}{1-\rho}.$$

Thus, (A.22) and $L_p - NL_d$ are equivalent to

$$
\begin{aligned}
&= \frac{\left(N\rho\sum_{i=0}^{N-2}\frac{N^i}{i!}\rho^i + \frac{N^N}{N!}\sum_{i=N}^{K} i\rho^i\right)(1-\rho)(1-\rho^{k+1})}{\left(\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i + \frac{N^N}{N!}\sum_{i=N}^{K}\rho^i\right)(1-\rho)(1-\rho^{k+1})} \\
&- \frac{N\rho\left(1-(k+1)\rho^k+k\rho^{k+1}\right)\left(\sum_{i=0}^{N-2}\frac{N^i}{i!}\rho^i + \frac{N^N}{N!}\frac{\rho^{N-1}-\rho^{K+1}}{1-\rho}\right)}{\left(\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i + \frac{N^N}{N!}\sum_{i=N}^{K}\rho^i\right)(1-\rho)(1-\rho^{k+1})} \\
&= \frac{\sum_{i=0}^{N-2}\frac{N^i}{i!}\rho^i\left(N\rho(1-\rho)(1-\rho^{k+1}) - N\rho\left(1-(k+1)\rho^k+k\rho^{k+1}\right)\right)}{\left(\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i + \frac{N^N}{N!}\sum_{i=N}^{K}\rho^i\right)(1-\rho)(1-\rho^{k+1})} \\
&+ \frac{\frac{N^N}{N!}}{\left(\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i + \frac{N^N}{N!}\sum_{i=N}^{K}\rho^i\right)(1-\rho)(1-\rho^{k+1})} \\
&\cdot \left(\frac{-(K+1)\rho^{K+1}+K\rho^{K+2}+N\rho^N-(N-1)\rho^{N+1}}{(1-\rho)^2}(1-\rho)(1-\rho^{k+1})\right. \\
&\left.-N\rho\left(1-(k+1)\rho^k+k\rho^{k+1}\right)\frac{\rho^{N-1}-\rho^{K+1}}{1-\rho}\right) \qquad \text{(A.23)} \\
&= \frac{\left(-\rho+(k+1)\rho^k-(k+1)\rho^{k+1}+\rho^{k+2}\right)N\rho\sum_{i=0}^{N-2}\frac{N^i}{i!}\rho^i}{(\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i + \frac{N^N}{N!}\sum_{i=N}^{K}\rho^i)(1-\rho)(1-\rho^{k+1})} \\
&+ \frac{(N^N/N!)\left[-(N-1)\rho^{N+1}-(K+1)\rho^{K+1}-N(k+1)\rho^{N+k+1}+N(k+1)\rho^{N+k}+(N-1)\rho^{N+k+2}\right]}{\left(\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i + \frac{N^N}{N!}\sum_{i=N}^{K}\rho^i\right)(1-\rho)^2(1-\rho^{k+1})} \\
&+ \frac{(N^N/N!)\left[(K+N)\rho^{K+2}+(K+1-Nk-N)\rho^{K+k+2}+(Nk-K)\rho^{K+k+3}\right]}{\left(\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i + \frac{N^N}{N!}\sum_{i=N}^{K}\rho^i\right)(1-\rho)^2(1-\rho^{k+1})}. \qquad \text{(A.24)}
\end{aligned}
$$

Equation (A.23) holds because

$$
\begin{aligned}
\sum_{i=N}^{K} i\rho^i &= \rho\frac{\partial}{\partial\rho}\left(\rho^N+\rho^{N+1}+\ldots+\rho^K\right) \\
&= \frac{-(K+1)\rho^{K+1}+K\rho^{K+2}+N\rho^N-(N-1)\rho^{N+1}}{(1-\rho)^2}. \qquad \text{(A.25)}
\end{aligned}
$$

The expression in (A.24), which is equivalent to $L_p - NL_d$, satisfies the following relations:

$$\begin{aligned}
&\frac{\left(-\rho+(k+1)\rho^k-(k+1)\rho^{k+1}+\rho^{k+2}\right)N\rho\sum_{i=0}^{N-2}\frac{N^i}{i!}\rho^i}{(\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i+\frac{N^N}{N!}\sum_{i=N}^{K}\rho^i)(1-\rho)(1-\rho^{k+1})}\\
&+\frac{\left[-(N-1)\rho^{N+1}-(K+1)\rho^{K+1}-N(k+1)\rho^{N+k+1}+N(k+1)\rho^{N+k}+(N-1)\rho^{N+k+2}\right]}{\left(\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i+\frac{N^N}{N!}\sum_{i=N}^{K}\rho^i\right)(1-\rho)^2(1-\rho^{k+1})}\cdot\frac{N^N}{N!}\\
&+\frac{(N^N/N!)\left[(K+N)\rho^{K+2}+(K+1-Nk-N)\rho^{K+k+2}+(Nk-K)\rho^{K+k+3}\right]}{\left(\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i+\frac{N^N}{N!}\sum_{i=N}^{K}\rho^i\right)(1-\rho)^2(1-\rho^{k+1})}\\
&<\frac{\left(-\rho+(k+1)\rho^k\right)N\rho\sum_{i=0}^{N-2}\frac{N^i}{i!}\rho^i}{\left(\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i+\frac{N^N}{N!}\sum_{i=N}^{K}\rho^i\right)(1-\rho)(1-\rho^{k+1})}\\
&+\frac{\frac{N^N}{N!}\frac{1}{1-\rho}\left(-(N-1)\rho^{N+1}-(K+1)\rho^{K+1}+N(k+1)\rho^{N+k}+(N-1)\rho^{N+k+2}+(K+N)\rho^{K+2}\right)}{\left(\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i+\frac{N^N}{N!}\sum_{i=N}^{K}\rho^i\right)(1-\rho)(1-\rho^{k+1})}
\end{aligned}\tag{A.26}$$

$$\begin{aligned}
&<\frac{\left(-\rho+(k+1)\rho^k\right)N\rho\sum_{i=0}^{N-2}\frac{N^i}{i!}\rho^i}{\left(\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i+\frac{N^N}{N!}\sum_{i=N}^{K}\rho^i\right)(1-\rho)(1-\rho^{k+1})}\\
&+\frac{\frac{N^N}{N!}\frac{1}{1-\rho}\left(-(N-1)\rho^{N+1}+(N-1)\rho^{Nk+1}+N(k+1)\rho^{N+k}+(N-1)\rho^{N+k+2}\right)}{\left(\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i+\frac{N^N}{N!}\sum_{i=N}^{K}\rho^i\right)(1-\rho)(1-\rho^{k+1})}
\end{aligned}\tag{A.27}$$

$$\begin{aligned}
&<\frac{\left(-\rho+(k+1)\rho^k\right)N\rho\sum_{i=0}^{N-2}\frac{N^i}{i!}\rho^i}{\left(\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i+\frac{N^N}{N!}\sum_{i=N}^{K}\rho^i\right)(1-\rho)(1-\rho^{k+1})}\\
&+\frac{\frac{1}{1-\rho}(-(N-1)\rho^{N+1}+N(k+3)\rho^{N+k})}{\left(\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i+\frac{N^N}{N!}\sum_{i=N}^{K}\rho^i\right)(1-\rho)(1-\rho^{k+1})}\frac{N^N}{N!}
\end{aligned}\tag{A.28}$$

$$<0.\tag{A.29}$$

The inequality (A.29) completes the proof of Lemma 10. Below we will explain how we obtain each of the inequalities above.

The inequality (A.26) holds because $-(k+1)\rho^{k+1}+\rho^{k+2} < -(k+1)\rho^{k+1}+\rho^{k+1} < 0$ for $\rho < 1$,

$$\begin{aligned}
&Nk \le K \le Nk+N-1, \text{ and} \\
&-N(k+1)\rho^{N+k+1}+(K+1-Nk-N)\rho^{K+k+2}+(Nk-K)\rho^{K+k+3} < 0.
\end{aligned}$$

The inequality (A.27) follows from the fact that $K \ge Nk$ and we have the following for $\rho < 1$:

$$\begin{aligned}
-(K+1)\rho^{K+1}+(K+N)\rho^{K+2} &< -(K+1)\rho^{K+1}+(K+N)\rho^{K+1} \\
&= (N-1)\rho^{K+1} \le (N-1)\rho^{Nk+1}
\end{aligned}$$

The inequality (A.28) is because

$$\begin{aligned}
&-(N-1)\rho^{N+1}+(N-1)\rho^{Nk+1}+N(k+1)\rho^{N+k}+(N-1)\rho^{N+k+2} \\
<&-(N-1)\rho^{N+1}+(N-1)\rho^{N+k}+N(k+1)\rho^{N+k}+(N-1)\rho^{N+k} \\
<&-(N-1)\rho^{N+1}+N(k+3)\rho^{N+k}.
\end{aligned}$$

The inequality (A.29) is due to the fact that $-\rho+(k+1)\rho^k < 0$ and $-(N-1)\rho^{N+1}+N(k+3)\rho^{N+k} < 0$ for $\rho < 1$ and $k > z_3$. Below we will prove these two inequalities. We already know from the proof of Lemma 9 that $g_4(z) \doteq (z+3)\rho^{z-1}$ is strictly decreasing in z for $\rho < 1$ and $z > -\frac{1}{\ln(\rho)} - 3$. Recall the definition of z_3 in (A.18). Because $z_3 \ge -\frac{1}{\ln(\rho)} - 3$ and $(z_3+3)\rho^{z_3-1} \le \frac{N-1}{N}$, we have the following for $k > z_3$:

$$(k+3)\rho^{k-1} < (N-1)/N.$$

Then,

$$\begin{aligned}
&-\rho+(k+1)\rho^k = \rho(-1+(k+1)\rho^{k-1}) < \rho\left(-1+(k+1)\frac{N-1}{N(k+3)}\right) < 0, \\
&-(N-1)\rho^{N+1}+N(k+3)\rho^{N+k} = (N-1)\rho^{N+1}\left(-1+\frac{N}{N-1}(k+3)\rho^{k-1}\right) < 0. \qquad \text{(A.30)}
\end{aligned}$$

This completes the proofs of the claims that $-\rho+(k+1)\rho^k < 0$ and $-(N-1)\rho^{N+1}+N(k+3)\rho^{N+k} < 0$ for $\rho < 1$ and $k > z_3$. □

We now use the result in Lemma 10 to prove Theorem 1-(b)-(i) under the condition ($\star\star$) in (2.12). For $\rho < 1$ and $k > z_3$,

$$W_p - W_d = \frac{L_p}{\lambda_{e,p}} - \frac{L_d}{\lambda_{e,d}} \tag{A.31}$$

$$< \frac{NL_d}{\lambda_{e,p}} - \frac{L_d}{\lambda_{e,d}} \tag{A.32}$$

$$< \frac{NL_d}{N\lambda_{e,d}} - \frac{L_d}{\lambda_{e,d}} \tag{A.33}$$

$$= 0. \tag{A.34}$$

The inequality (A.31) follows from Little's Law. The inequality (A.32) holds because $L_p < NL_d$ by Lemma 10. Recall from Lemma 8 that $\lambda_{e,p} > N\lambda_{e,d}$ regardless of the value of ρ. This implies the inequality (A.33). The definition of k and (A.34) complete the proof of Theorem 1-(b)-(i) under the condition ($\star\star$).

We now show Theorem 1-(b)-(ii) under the condition ($\star\star$) in (2.12). Recall that

$$SW_d = \lambda_{e,d}N(R - cW_d) \qquad \text{and} \qquad SW_p = \lambda_{e,p}(R - cW_p).$$

Because $\lambda_{e,d}N < \lambda_{e,p}$ by Lemma 8 and $W_d > W_p$ by part (b)-(i), we have $SW_d < SW_p$. This completes the proof of Theorem 1-(b)-(ii) under the condition ($\star\star$) in (2.12). □

A.3 Proof of Proposition 1

The proof follows from Lemma 8 in Appendix A.2 of the Electronic Companion. □

A.4 Proof of Proposition 2

Denote by X_s the number of customers in the SQ system in the steady-state, and let X_p be the corresponding figure in the pooled system. Note that the SQ system is the same as the M/M/1/K system described in Lemma 7. To show our claim, we will use the standard likelihood comparison technique. Let $\gamma_s(m+1)$ be the transition rate from state $m+1$ to m in the SQ system, $\gamma_p(m+1)$ be the transition rate from state $m+1$ to m in the pooled system, for any $m = 0, 1, \ldots K-1$. Because $\gamma_p(m+1) \leq \gamma_s(m+1)$ for each

m, in the steady-state, we have

$$\mathbb{P}(X_p = m)N\lambda \leq \mathbb{P}(X_p = m+1)\gamma_s(m+1) \quad \text{and} \quad \mathbb{P}(X_s = m)N\lambda = \mathbb{P}(X_s = m+1)\gamma_s(m+1).$$

Thus, we have

$$\frac{\mathbb{P}(X_p = m+1)}{\mathbb{P}(X_p = m)} \geq \frac{N\lambda}{\gamma_s(m+1)} = \frac{\mathbb{P}(X_s = m+1)}{\mathbb{P}(X_s = m)}. \tag{A.35}$$

Using this, we now show that $L_p \doteq \mathbb{E}(X_p) \geq L_s \doteq \mathbb{E}(X_s)$. Note that (A.35) implies that $\frac{\mathbb{P}(X_p=j)}{\mathbb{P}(X_p=i)} \geq \frac{\mathbb{P}(X_s=j)}{\mathbb{P}(X_s=i)}$ for all $i \leq j$, $i, j \in \{0, 1, \ldots, K\}$, which is equivalent to

$$\mathbb{P}(X_p = j)\mathbb{P}(X_s = i) \geq \mathbb{P}(X_p = i)\mathbb{P}(X_s = j). \tag{A.36}$$

The summation on both sides of (A.36) over i from 0 to j gives

$$\mathbb{P}(X_p = j)\mathbb{P}(X_s \leq j) \geq \mathbb{P}(X_p \leq j)\mathbb{P}(X_s = j). \tag{A.37}$$

Similarly, the summation on both sides of (A.36) over j from $i+1$ to K results in

$$\mathbb{P}(X_p \geq i+1)\mathbb{P}(X_s = i) \geq \mathbb{P}(X_p = i)\mathbb{P}(X_s \geq i+1). \tag{A.38}$$

Combining (A.37) and (A.38) and letting $i = j = a$, we have

$$\frac{\mathbb{P}(X_p \geq a+1)}{\mathbb{P}(X_s \geq a+1)} \geq \frac{\mathbb{P}(X_p = a)}{\mathbb{P}(X_s = a)} \geq \frac{\mathbb{P}(X_p \leq a)}{\mathbb{P}(X_s \leq a)}. \tag{A.39}$$

Thus, $\mathbb{P}(X_p \leq a) \leq \mathbb{P}(X_s \leq a)$ for any $a \in \{0, 1, 2, \ldots, K\}$, hence

$$L_p = \mathbb{E}(X_p) = \sum_{i=0}^{K}(1 - \mathbb{P}(X_p \leq i)) \geq \sum_{i=0}^{K}(1 - \mathbb{P}(X_s \leq i)) = \mathbb{E}(X_s) = L_s. \tag{A.40}$$

Recall $\lambda_{e,p}$ and $\lambda_{e,s}$ from (A.4) and (A.9), respectively. Then,

$$\lambda_{e,p} = \left(1 - \frac{\frac{N^N}{N!}\rho^K}{\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i + \frac{N^N}{N!}\sum_{i=N}^{K}\rho^i}\right)N\lambda < \left(1 - \frac{\rho^K}{\sum_{i=0}^{K}\rho^i}\right)N\lambda = \lambda_{e,s}. \tag{A.41}$$

By (A.40), (A.41) and Little's Law,

$$W_p = \frac{L_p}{\lambda_{e,p}} \geq \frac{L_s}{\lambda_{e,p}} > \frac{L_s}{\lambda_{e,s}} = W_s.$$

Recall the definition of social welfare from (2.3):

$$SW_p = \lambda_{e,p}(R - cW_p) \qquad \text{and} \qquad SW_s = \lambda_{e,s}(R - cW_s).$$

Because $\lambda_{e,s} > \lambda_{e,p}$ and $W_s < W_p$, $SW_s > SW_p$. This completes the proof of the claim. □

A.5 Proof of Proposition 3 and the Statement and the Proof of Proposition 20

Recall Lemmas 1, 2 and 7. To prove Proposition 3, we shall define some constants. Let

$$\eta \doteq (z_1 + 1)/\mu$$

where

$$z_1 \doteq \inf\left\{ z \in \mathbb{R} : z > \frac{1}{\ln(\rho)} - 1, \max\left\{ \frac{(z+1)N\rho}{\rho^{z+1}-1} - \frac{N-1}{4(\rho-1)^2}, \frac{(z+1)N}{(N-1)\rho^z} - \frac{1}{2} \right\} < 0 \right\}. \quad \text{(A.42)}$$

Proof of Proposition 3: The outline of our proof is as follows. First, we will state and prove Lemma 11 that shows the existence of the constant z_1 defined in (A.42). Then, we will show that if $\rho > 1$ and $k > z_1$,

$$SW_d - SW_s > c\frac{(N-1)}{(\rho-1)4}. \quad \text{(A.43)}$$

This and the fact that $\frac{R}{c} > \frac{z_1+1}{\mu}$ implies $k > z_1$ complete the proof of social welfare claim in Proposition 3-(b). Recall the social welfare from (2.3). Because $SW_d > SW_s$ by Proposition 3-(b), and $\lambda_{e,d}N < \lambda_{e,s}$ by Proposition 1 and the proof of Proposition 2, $W_d < W_s$. Thus, the average sojourn time claim in part (a) follows.

Lemma 11. *For $\rho > 1$, the constant z_1 defined in $(A.42)$ exists and $z_1 \in [1, \infty)$.*

Proof of Lemma 11: Define $g_1(z) \doteq \frac{N\rho(z+1)}{\rho^{z+1}-1}$ and $g_2(z) \doteq \frac{N(z+1)}{N-1}\rho^{-z}$. Then, note that the definition in (A.42) is equivalent to $z_1 \doteq \inf\{z \in \mathbb{R} : g_1(z) < (N-1)/(4(\rho-1)^2), g_2(z) < 1/2 \text{ and } z > 1/\ln(\rho) - 1\}$.

Both $g_1(\cdot)$ and $g_2(\cdot)$ are strictly decreasing when $\rho > 1$ and $z > \frac{1}{\ln(\rho)} - 1$ because

$$g_1'(z) = N\rho\frac{\rho^{z+1} - 1 - (z+1)\rho^{z+1}\ln(\rho)}{(\rho^{z+1}-1)^2} = N\rho\frac{\rho^{z+1}\left(1-(z+1)\ln(\rho)\right) - 1}{(\rho^{z+1}-1)^2} < 0, \text{ and} \tag{A.44}$$

$$g_2'(z) = \frac{N}{N-1}\frac{\rho^z - (z+1)\rho^z\ln(\rho)}{\rho^{2z}} = \frac{N}{N-1}\frac{\rho^z\left(1-(z+1)\ln(\rho)\right)}{\rho^{2z}} < 0, \tag{A.45}$$

for $\rho > 1$ and $z > \frac{1}{ln(\rho)} - 1$. In addition, we have

$$\lim_{z\to\infty} g_1(z) = 0 \qquad \text{and} \qquad \lim_{z\to\infty} g_2(z) = 0. \tag{A.46}$$

It follows from (A.44) through (A.46) that z_1 exists and it is finite. We now show that $z_1 \geq 1$ for $\rho > 1$. Suppose for a contradiction that $z_1 < 1$. Because $g_1(z)$ and $g_2(z)$ are strictly decreasing for $\rho > 1$ and $z > \frac{1}{\ln(\rho)} - 1$, and $z_1 \geq \frac{1}{\ln(\rho)} - 1$ by definition of z_1, at $z = 1$ the following inequalities must hold: $g_1(1) < \frac{N-1}{4(\rho-1)^2}$ and $g_2(1) < \frac{1}{2}$. Note that

$$g_2(1) = \frac{2N}{(N-1)\rho} < \frac{1}{2} \Leftrightarrow \rho > \frac{4N}{N-1}, \tag{A.47}$$

which implies that $\rho > 4$. Observe also that

$$g_1(1) = \frac{2N\rho}{\rho^2-1} < \frac{N-1}{4(\rho-1)^2} \Leftrightarrow \left(\frac{(\rho-1)2N\rho}{\rho+1} - \frac{N-1}{4}\right)\frac{1}{(\rho-1)^2} < 0. \tag{A.48}$$

But, for $\rho > 4$,

$$\frac{(\rho-1)2N\rho}{\rho+1} - \frac{N-1}{4} > \frac{24N}{5} - \frac{N-1}{4} > 0,$$

which contradicts (A.48). Thus, $z_1 \geq 1$. □

We begin with the proof of Proposition 3-(b). Recall from (A.8) that the social welfare in the dedicated system is

$$\begin{aligned} SW_d &= \left(\frac{1-\rho^k}{1-\rho^{k+1}}\right)RN\lambda - \left(\frac{\rho}{1-\rho} - \frac{(k+1)\rho^{k+1}}{1-\rho^{k+1}}\right)Nc \\ &= RN\lambda - \frac{1-\rho}{1-\rho^{k+1}}\left(N\lambda R\rho^k + Nc\frac{\rho-(1+k)\rho^{k+1}+k\rho^{k+2}}{(\rho-1)^2}\right). \end{aligned} \tag{A.49}$$

Recall also from (A.12) that

$$SW_s = \left(\frac{1-\rho^K}{1-\rho^{K+1}}\right) RN\lambda - \left(\frac{\rho}{1-\rho} - \frac{(K+1)\rho^{K+1}}{1-\rho^{K+1}}\right) c$$
$$= RN\lambda - \frac{1-\rho}{1-\rho^{K+1}}\left(N\lambda R\rho^K + c\frac{\rho-(1+K)\rho^{K+1}+K\rho^{K+2}}{(\rho-1)^2}\right). \tag{A.50}$$

From (A.49) and (A.50), it follows that $SW_d - SW_s > c\frac{(N-1)}{(\rho-1)4}$ if and only if

$$\frac{1}{\rho^{k+1}-1}\left(RN\lambda\rho^k + Nc\frac{\rho-(1+k)\rho^{k+1}+k\rho^{k+2}}{(\rho-1)^2}\right)$$
$$- \frac{1}{\rho^{K+1}-1}\left(RN\lambda\rho^K + c\frac{\rho-(1+K)\rho^{K+1}+K\rho^{K+2}}{(\rho-1)^2}\right) < -c\frac{(N-1)}{4(\rho-1)^2}.$$

Note that the left hand side of the above inequality is equivalent to

$$RN\lambda\left(\frac{\rho^k}{\rho^{k+1}-1} - \frac{\rho^K}{\rho^{K+1}-1}\right) - c\left(\frac{\rho-(1+K)\rho^{K+1}+K\rho^{K+2}}{(\rho-1)^2(\rho^{K+1}-1)} - N\frac{\rho-(1+k)\rho^{k+1}+k\rho^{k+2}}{(\rho-1)^2(\rho^{k+1}-1)}\right).$$

Rearranging this, we conclude that $SW_d - SW_s > c\frac{(N-1)}{(\rho-1)4}$ if and only if

$$\underbrace{N\lambda\frac{R}{c}\left(\frac{\rho^k}{\rho^{k+1}-1} - \frac{\rho^K}{\rho^{K+1}-1}\right)}_{\text{First Term}} - \underbrace{\left(\frac{\rho-(1+K)\rho^{K+1}+K\rho^{K+2}}{(\rho-1)^2(\rho^{K+1}-1)} - N\frac{\rho-(1+k)\rho^{k+1}+k\rho^{k+2}}{(\rho-1)^2(\rho^{k+1}-1)}\right)}_{\text{Second Term}}$$
$$< -\frac{(N-1)}{4(\rho-1)^2} \tag{A.51}$$

We claim and show below that the first term in (A.51) is bounded above by $(N-1)/(4(\rho-1)^2)$ if $\rho > 1$ and $k > z_1$. Next we will show that the second term in (A.51) is bounded below by $(N-1)/(2(\rho-1)^2)$ if $\rho > 1$ and $k > z_1$. These two results imply that if $\rho > 1$ and $k > z_1$, we have (A.51). From this, Proposition 3-(b) follows.

We now show our claim that if $\rho > 1$ and $k > z_1$, the first term in (A.51) is bounded above by $(N-1)/(4(\rho-1)^2)$. Suppose that $\rho > 1$. Then, the first term in (A.51) satisfies the following inequality:

$$N\lambda\frac{R}{c}\left(\frac{\rho^k}{\rho^{k+1}-1}-\frac{\rho^K}{\rho^{K+1}-1}\right) < N\lambda\frac{k+1}{\mu}\left(\frac{\rho^k}{\rho^{k+1}-1}-\frac{\rho^K}{\rho^{K+1}-1}\right) \tag{A.52}$$

$$< N\lambda\frac{k+1}{\mu}\left(\frac{\rho^k}{\rho^{k+1}-1}-\frac{\rho^{Nk+N}}{\rho^{Nk+N+1}-1}\right) \tag{A.53}$$

$$= N\rho(k+1)\frac{\rho^{Nk+N}-\rho^k}{(\rho^{k+1}-1)(\rho^{Nk+N+1}-1)} \tag{A.54}$$

$$< \frac{N\rho(k+1)}{(\rho^{k+1}-1)}. \tag{A.55}$$

The inequality (A.52) above follows from the definition of k in (2.5) and the fact that the first term in (A.51) is positive. Note from the definitions of k and K in (2.5) and (2.6) that we also have $K < Nk+N$. Then, the inequality (A.53) follows because $K < Nk+N$ and $\frac{\rho^K}{\rho^{K+1}-1}$ is decreasing in K for $\rho > 1$. The inequality (A.55) is because $\rho^{Nk+N} - \rho^k < \rho^{Nk+N+1} - 1$ for $\rho > 1$.

We already know from the proof of Lemma 11 that $g_1(z) \doteq \frac{(z+1)N\rho}{(\rho^{z+1}-1)}$ is strictly decreasing in z for $\rho > 1$ and $z > \frac{1}{\ln(\rho)} - 1$. Recall the definition of z_1 in (A.42). Because $z_1 \geq \frac{1}{\ln(\rho)} - 1$ and $\frac{(z_1+1)N\rho}{(\rho^{z_1+1}-1)} \leq \frac{N-1}{4(\rho-1)^2}$, we have the following for $k > z_1$:

$$\frac{N\rho(k+1)}{(\rho^{k+1}-1)} < \frac{N-1}{4(\rho-1)^2},$$

which together with (A.55) implies that

$$N\lambda\frac{R}{c}\left(\frac{\rho^k}{\rho^{k+1}-1}-\frac{\rho^K}{\rho^{K+1}-1}\right) < \frac{N-1}{4(\rho-1)^2}. \tag{A.56}$$

This completes our arguments for the proof of our first claim above that the first term in (A.51) is bounded above by $(N-1)/(4(\rho-1)^2)$ if $\rho > 1$ and $k > z_1$.

We now show our claim that if $\rho > 1$ and $k > z_1$, the second term in (A.51) is bounded below by $(N-1)/(2(\rho-1)^2)$. To prove this claim, we first prove that $h(\rho, x) \doteq \frac{\rho-(1+x)\rho^{x+1}+x\rho^{x+2}}{(\rho-1)^2(\rho^{x+1}-1)}$ is increasing in

x for $\rho > 1$. Note that

$$h(\rho, x) = \frac{1}{\rho - 1} \frac{\rho - (1+x)\rho^{x+1} + x\rho^{x+2}}{(\rho-1)(\rho^{x+1}-1)} = \frac{1}{\rho-1} \frac{\rho - (1+x)\rho^{x+1} + x\rho^{x+2}}{(1-\rho)(1-\rho^{x+1})}$$
$$= \frac{1}{\rho-1}\left(\frac{\rho}{1-\rho} - \frac{(x+1)\rho^{x+1}}{1-\rho^{x+1}}\right).$$

Then,

$$\frac{\partial h(\rho, x)}{\partial x} = -\frac{\rho^{x+1}}{(\rho-1)(1-\rho^{x+1})^2}(1 - \rho^{x+1} + (x+1)\ln(\rho)). \tag{A.57}$$

Let $v(\rho, x) \doteq 1 - \rho^{x+1} + (x+1)\ln(\rho)$ and observe from (A.57) that $v(\rho, x)$ and $\frac{\partial h(\rho,x)}{\partial x}$ have opposite signs for $\rho > 1$. Note that $v(1, x) = 0$ for any x, and for $\rho > 1$ and $x \geq 0$,

$$\frac{\partial v(\rho, x)}{\partial \rho} = -(x+1)\rho^x + (x+1)\frac{1}{\rho} = (x+1)\left(-\rho^x + \frac{1}{\rho}\right) < 0.$$

This immediately implies that $\frac{\partial h(\rho,x)}{\partial x} > 0$ for $\rho > 1$ and $x \geq 0$. Based on this, for $\rho > 1$, the second term in (A.51) satisfies

$$\begin{aligned}
&\frac{\rho - (1+K)\rho^{K+1} + K\rho^{K+2}}{(\rho-1)^2(\rho^{K+1}-1)} - N\frac{\rho - (1+k)\rho^{k+1} + k\rho^{k+2}}{(\rho-1)^2(\rho^{k+1}-1)} \\
\geq &\frac{\rho - (1+Nk)\rho^{Nk+1} + Nk\rho^{Nk+2}}{(\rho-1)^2(\rho^{Nk+1}-1)} - N\frac{\rho - (1+k)\rho^{k+1} + k\rho^{k+2}}{(\rho-1)^2(\rho^{k+1}-1)} && \text{(A.58)} \\
= &\frac{(N-1)\rho^{(N+1)k+2} - N(k+1)\rho^{Nk+2} + (Nk+1)\rho^{Nk+1} + \rho^{k+2} - \rho + N(\rho - (1+k)\rho^{k+1} + k\rho^{k+2})}{(\rho^{Nk+1}-1)(\rho^{k+1}-1)} \\
&\cdot \frac{1}{(\rho-1)^2} \\
> &\frac{1}{(\rho-1)^2}\frac{(N-1)\rho^{(N+1)k+2} - N(k+1)\rho^{Nk+2}}{(\rho^{Nk+1}-1)(\rho^{k+1}-1)} && \text{(A.59)} \\
= &\frac{1}{(\rho-1)^2}\frac{(N-1)\rho^{(N+1)k+2} - N(k+1)\rho^{Nk+2}}{\rho^{(N+1)k+2} - \rho^{Nk+1} - \rho^{k+1} + 1} \\
= &\frac{N-1}{(\rho-1)^2}\frac{1 - \frac{N(k+1)}{N-1}\rho^{-k}}{1 - \rho^{-k-1} - \rho^{-Nk-1} + \rho^{-(Nk+k+2)}}. && \text{(A.60)}
\end{aligned}$$

We have (A.58) because $K \geq Nk$ and as we showed earlier, $h(\rho, x) = \frac{\rho-(1+x)\rho^{x+1}+x\rho^{x+2}}{(\rho-1)^2(\rho^{x+1}-1)}$ is increasing in x for $\rho > 1$. The inequality (A.59) holds because $W_d > 0$ in the first line of (A.7) implies that $\rho - (1 + k)\rho^{k+1} + k\rho^{k+2} > 0$. We already know from the proof of Lemma 11 that $g_2(z) = \frac{(z+1)N}{N-1}\rho^{-z}$ is strictly

decreasing in z if $z > \frac{1}{\ln(\rho)} - 1$ and $\rho > 1$, and $\lim_{z\to\infty} g_2(z) = 0$. This, the definition of z_1 in (A.42) and the fact that $(1 - \rho^{-k-1} - \rho^{-Nk-1} + \rho^{-(Nk+k+2)}) \in (0, 1)$ for $\rho > 1$ imply that

$$\frac{N-1}{(\rho-1)^2} \frac{1 - \frac{N(k+1)}{N-1}\rho^{-k}}{1 - \rho^{-k-1} - \rho^{-Nk-1} + \rho^{-(Nk+k+2)}} > \frac{N-1}{2(\rho-1)^2} \quad \text{for } k > z_1,\ \rho > 1. \tag{A.61}$$

Combining this and (A.60), it follows that if $\rho > 1$ and $k > z_1$,

$$\frac{\rho - (1+K)\rho^{K+1} + K\rho^{K+2}}{(\rho-1)^2(\rho^{K+1}-1)} - N\frac{\rho - (1+k)\rho^{k+1} + k\rho^{k+2}}{(\rho-1)^2(\rho^{k+1}-1)} > \frac{N-1}{2(\rho-1)^2}. \tag{A.62}$$

This completes the proof of our claim that if $\rho > 1$ and $k > z_1$, the second term in (A.51) is bounded below by $(N-1)/(2(\rho-1)^2)$, and hence completes our arguments for the proof of Proposition 3-(b).

Now we prove Proposition 3-(a). Recall that

$$SW_d = \lambda_{e,d}N(R - cW_d) \quad \text{and} \quad SW_s = \lambda_{e,s}(R - cW_s).$$

By Proposition 1, $\lambda_{e,d}N < \lambda_{e,p}$ and we already know from the proof of Proposition 2 that $\lambda_{e,p} < \lambda_{e,s}$. These imply

$$\lambda_{e,d}N < \lambda_{e,s}. \tag{A.63}$$

From this and Proposition 3-(b), the claim immediately follows. □

Proposition 20. *The dedicated system results in* **(i)** *strictly larger average sojourn time and* **(ii)** *strictly smaller social welfare than the SQ system, i.e., $W_d > W_s$ and $SW_d < SW_s$, respectively, if*

$$\rho < 1 \quad \textit{and} \quad R/c > \widetilde{\eta} \doteq (z_2 + 1)/\mu, \tag{A.64}$$

where

$$z_2 \doteq \inf\{z \in \mathbb{R} : z > -1/\ln(\rho), z\rho^z < (N-1)/N\}. \tag{A.65}$$

Proof of Proposition 20: We first prove a lemma which will be used in the remainder of the proof.

Lemma 12. *For $\rho < 1$, the constant z_2 defined in $(A.65)$ exists and it is finite.*

Proof of Lemma 12: Define $g_3(z) \doteq z\rho^z$. Then, note that the definition in (A.65) is equivalent to $z_2 \doteq \inf\{z \in \mathbb{R} : g_3(z) < (N-1)/N, \text{ and } z > -1/\ln(\rho)\}$. The function $g_3(\cdot)$ is strictly decreasing when $\rho < 1$ and $z > -\frac{1}{\ln(\rho)}$ because

$$g_3'(z) = \rho^z + z\rho^z \ln(\rho) = \rho^z(1 + z\ln(\rho)) < 0, \tag{A.66}$$

for $\rho < 1$ and $z > -\frac{1}{\ln(\rho)}$. In addition, we have

$$\lim_{z\to\infty} g_3(z) = 0. \tag{A.67}$$

It follows from (A.66) and (A.67) that z_2 exists and it is finite. □

Recall from (A.7) that when $\rho < 1$, the average sojourn time in the dedicated system is

$$W_d = \frac{\rho - (k+1)\rho^{k+1} + k\rho^{k+2}}{\lambda(\rho^k - 1)(\rho - 1)}.$$

Recall from (A.11) that when $\rho < 1$, the average sojourn time in the SQ system is

$$W_s = \frac{1}{N\lambda(\rho - 1)} \frac{K\rho^{K+2} - (K+1)\rho^{K+1} + \rho}{\rho^K - 1}.$$

If $\rho < 1$ and $k > z_2$,

$$
\begin{aligned}
& W_d - W_s \\
=& \frac{\rho - (k+1)\rho^{k+1} + k\rho^{k+2}}{\lambda(\rho^k - 1)(\rho - 1)} - \frac{1}{N\lambda(\rho-1)} \frac{K\rho^{K+2} - (K+1)\rho^{K+1} + \rho}{\rho^K - 1} \\
\geq& \frac{\rho - (k+1)\rho^{k+1} + k\rho^{k+2}}{\lambda(\rho^k - 1)(\rho - 1)} - \frac{1}{N\lambda(\rho-1)} \frac{(Nk+N-1)\rho^{Nk+N+1} - (Nk+N)\rho^{Nk+N} + \rho}{\rho^{Nk+N-1} - 1} \qquad \text{(A.68)} \\
=& \Big((N-1)(\rho^{Nk+k+N+1} - \rho^{Nk+N+1} - \rho^{k+1} + \rho) + Nk(\rho^{Nk+N} - \rho^{Nk+N+1} + \rho^{k+2} - \rho^{k+1}) \Big) \\
& \frac{1}{(1-\rho)(1-\rho^k)(1-\rho^{Nk+N-1})N\lambda} \\
=& \Big((N-1)(\rho - \rho^{Nk+N+1})(1-\rho^k) - Nk(\rho^{k+1} - \rho^{Nk+N})(1-\rho) \Big) \frac{1}{(1-\rho)(1-\rho^k)(1-\rho^{Nk+N-1})N\lambda} \\
>& \Big((N-1)(\rho - \rho^{Nk+N+1}) - Nk(\rho^{k+1} - \rho^{Nk+N}) \Big) \frac{1}{(1-\rho^k)(1-\rho^{Nk+N-1})N\lambda} \\
>& \Big((N-1)\rho^{-k} - Nk \Big) \frac{\rho^{k+1} - \rho^{Nk+N}}{(1-\rho^k)(1-\rho^{Nk+N-1})N\lambda} \qquad \text{(A.69)} \\
>& 0, \qquad \text{(A.70)}
\end{aligned}
$$

which proves the claim in part (i).

We now explain why the inequalities in (A.68) and (A.70) hold. Note that W_s is increasing in K because

$$
\begin{aligned}
& \frac{\partial(\frac{1}{N\lambda(\rho-1)} \frac{K\rho^{K+2} - (K+1)\rho^{K+1} + \rho}{\rho^K - 1})}{\partial K} \\
=& \frac{1}{N\lambda(\rho-1)} \frac{-\rho^{K+2} - K\rho^{K+2}\ln(\rho) + \rho^{K+1} + K\rho^{K+1}\ln(\rho) + \rho^{2K+2} - \rho^{2K+1}}{(\rho^K - 1)^2} \\
=& \frac{1}{N\lambda(\rho-1)} \cdot \frac{(\rho-1)\rho^{K+1}(\rho^K - K\ln(\rho) - 1)}{(\rho^K - 1)^2} \\
=& \frac{1}{N\lambda} \cdot \frac{\rho^{K+1}(\rho^K - K\ln(\rho) - 1)}{(\rho^K - 1)^2} \\
\geq& 0. \qquad \text{(A.71)}
\end{aligned}
$$

(Here, (A.71) follows from the fact that $\rho^K - K\ln(\rho) - 1 \geq 0$, which is because $\frac{\partial(\rho^K - K\ln(\rho) - 1)}{\partial \rho} = K\rho^{K-1} - \frac{K}{\rho} = \frac{K}{\rho}(\rho^K - 1)$ and thus $\rho^K - K\ln(\rho) - 1$ achieves the minimum, which is 0, at $\rho = 1$.) Then, (A.68) follows from the fact that $K \leq Nk + N - 1$ and W_s is increasing in K, as shown above.

The reason for (A.70) is as follows. We already know from the proof of Lemma 12 that $g_3(z) \doteq z\rho^z$ is strictly decreasing in z for $\rho < 1$ and $z > -\frac{1}{\ln(\rho)}$. Recall the definition of z_2 in (A.65). Because $z_2 \geq -\frac{1}{\ln(\rho)}$ and $z_2\rho^{z_2} \leq \frac{N-1}{N}$, we have $k\rho^k < \frac{N-1}{N}$ for $\rho < 1$ and $k > z_2$.

We now prove the claim in part (ii). Recall from (2.3) that

$$SW_d = \lambda_{e,d}N(R - cW_d) \quad \text{and} \quad SW_s = \lambda_{e,s}(R - cW_s).$$

Since $\lambda_{e,d}N < \lambda_{e,s}$ by Proposition 1 and the proof of Proposition 2, part (i) implies part (ii). □

A.6 Statement and Proof of Lemma 13

Lemma 13. *Consider any fixed service rate μ.* **(a)** *As $\rho \to \infty$, W_d and W_p satisfy the following relations:*

$$\lim_{\rho\to\infty} W_d(\rho) = \frac{\lfloor R\mu/c \rfloor}{\mu} \leq \lim_{\rho\to\infty} W_p(\rho) = \frac{\lfloor RN\mu/c \rfloor}{N\mu}, \tag{A.72}$$

$$\lim_{\rho\to\infty} W_d'(\rho) = \lim_{\rho\to\infty} W_p'(\rho) = 0. \tag{A.73}$$

(b) *As $\rho \to 1$, W_d and W_p satisfy the following relations:*

$$\lim_{\rho\to 1} [W_p(\rho) - W_d(\rho)] > 0 \tag{A.74}$$

if $\lfloor RN\mu/c \rfloor - N\lfloor R\mu/c \rfloor > \left(\sum_{i=0}^{N-2} N^i/i!\right) / \left(N^N/N!\right)$, and for $R/c > 10/\mu$,

$$\lim_{\rho\to 1} \left[W_p'(\rho) - W_d'(\rho)\right] > \left(\lfloor R\mu/c \rfloor\right)^2 \gamma(N)/\mu^2 > 0, \tag{A.75}$$

where $\gamma(N) > 0$ is a linear function of N and does not depend on other parameters.

(c) *As $\rho \to 0$, W_d and W_p satisfy the following relations:*

$$\lim_{\rho\to 0} W_d(\rho) = \lim_{\rho\to 0} W_p(\rho) = \frac{1}{\mu} \quad \textit{and} \quad \lim_{\rho\to 0} W_d'(\rho) \geq \lim_{\rho\to 0} W_p'(\rho) = 0. \tag{A.76}$$

Note that for a fixed μ, the limits $\rho \to 0$, $\rho \to 1$ and $\rho \to \infty$ are equivalent to $\lambda \to 0$, $\lambda \to \mu$ and $\lambda \to \infty$, respectively. In this proof, we will use these equivalent limits.

We first identify $W_d'(\lambda)$ and $W_p'(\lambda)$ and we will use those expressions to prove parts (a) through (c) of Proposition 13. Recall from (A.7) that the average sojourn time in the dedicated system is

$$W_d(\lambda) = \frac{\rho - (k+1)\rho^{k+1} + k\rho^{k+2}}{\lambda(\rho^k - 1)(\rho - 1)} = \frac{1}{\mu}\frac{\left(1 - (k+1)\rho^k + k\rho^{k+1}\right)}{(\rho^k - 1)(\rho - 1)}, \quad \rho \neq 1. \tag{A.77}$$

Thus, we have

$$\begin{aligned} W_d'(\lambda) =& \frac{1}{\mu^2}\frac{(-(k+1)k\rho^{k-1} + k(k+1)\rho^k)(\rho^k - 1)(\rho - 1) - (1 - (k+1)\rho^k + k\rho^{k+1})}{(\rho - 1)^2(\rho^k - 1)^2} \\ & \cdot ((k+1)\rho^k - k\rho^{k-1} - 1) \\ =& \frac{1}{\mu^2}\frac{\rho^{2k} - k^2\rho^{k+1} + (2k^2 - 2)\rho^k - k^2\rho^{k-1} + 1}{(\rho - 1)^2(\rho^k - 1)^2}. \end{aligned} \tag{A.78}$$

Recall (2.7). When $K = N$,

$$W_p(\lambda) = \frac{1}{\mu}, \tag{A.79}$$

and when $K > N$,

$$W_p(\lambda) = \frac{\sum_{i=0}^{N-1}\frac{N^i}{i!}i\rho^i + \frac{N^N}{N!}\sum_{i=N}^{K} i\rho^i}{\left(\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i + \frac{N^N}{N!}\sum_{i=N}^{K-1}\rho^i\right)N\lambda} \tag{A.80}$$

$$= \frac{\sum_{i=0}^{N}\frac{N^i}{i!}i\rho^i + \frac{N^N}{N!}\sum_{i=N+1}^{K} i\rho^i}{\left(\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i + \frac{N^N}{N!}\sum_{i=N}^{K-1}\rho^i\right)N\lambda}$$

$$= \frac{1}{(N\mu)\rho}\frac{N\rho\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i + \frac{N^N}{N!}\sum_{i=N+1}^{K} i\rho^i}{\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i + \frac{N^N}{N!}\sum_{i=N}^{K-1}\rho^i} \tag{A.81}$$

$$= \frac{1}{N\mu}\frac{N\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i + \frac{N^N}{N!}\sum_{i=N+1}^{K} i\rho^{i-1}}{\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i + \frac{N^N}{N!}\sum_{i=N}^{K-1}\rho^i}. \tag{A.82}$$

The equation (A.81) is due to the fact that

$$\sum_{i=0}^{N}\frac{N^i}{i!}i\rho^i = \sum_{i=1}^{N}\frac{N^i}{(i-1)!}\rho^i = N\rho\sum_{i=1}^{N}\frac{N^{i-1}}{(i-1)!}\rho^{i-1} = N\rho\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i.$$

Based on (A.79) and (A.82),

$$W_p'(\lambda) = 0 \tag{A.83}$$

for $K = N$, and when $K > N$, we have

$$\begin{aligned} W_p'(\lambda) =& \left[\frac{\partial}{\partial \rho}\left(\frac{N\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i + \frac{N^N}{N!}\sum_{i=N+1}^{K} i\rho^{i-1}}{\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i + \frac{N^N}{N!}\sum_{i=N}^{K-1}\rho^i}\right)\right]\frac{1}{N\mu^2} \\ =& \frac{\left(N\sum_{i=1}^{N-1}\frac{N^i}{i!}i\rho^{i-1} + \frac{N^N}{N!}\sum_{i=N+1}^{K} i(i-1)\rho^{i-2}\right)\left(\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i + \frac{N^N}{N!}\sum_{i=N}^{K-1}\rho^i\right)}{\left(\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i + \frac{N^N}{N!}\sum_{i=N}^{K-1}\rho^i\right)^2 N\mu^2} \\ & - \frac{\left(N\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i + \frac{N^N}{N!}\sum_{i=N+1}^{K} i\rho^{i-1}\right)\left(\sum_{i=1}^{N-1}\frac{N^i}{i!}i\rho^{i-1} + \frac{N^N}{N!}\sum_{i=N}^{K-1} i\rho^{i-1}\right)}{\left(\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i + \frac{N^N}{N!}\sum_{i=N}^{K-1}\rho^i\right)^2 N\mu^2}. \end{aligned} \tag{A.84}$$

Proof of Part (a): From (A.77), we have

$$\lim_{\lambda\to\infty} W_d(\lambda) = \lim_{\lambda\to\infty}\frac{1}{\mu}\frac{\left(1-(k+1)\rho^k + k\rho^{k+1}\right)}{(\rho^k-1)(\rho-1)} = \frac{k}{\mu} \tag{A.85}$$

because the leading term in the above numerator is $k\rho^{k+1}$ and the leading term in the above denominator is $\mu\rho^{k+1}$ as $\lambda \to \infty$.

From (A.79), when $K = N$,

$$\lim_{\lambda\to\infty} W_p(\lambda) = \frac{1}{\mu} = \frac{K}{N\mu}. \tag{A.86}$$

From (A.82), it follows that for $K > N$,

$$\begin{aligned} \lim_{\lambda\to\infty} W_p(\lambda) =& \frac{1}{N\mu}\lim_{\lambda\to\infty}\frac{N\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i + \frac{N^N}{N!}\sum_{i=N+1}^{K} i\rho^{i-1}}{\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i + \frac{N^N}{N!}\sum_{i=N}^{K-1}\rho^i} \\ =& \frac{1}{N\mu}\lim_{\lambda\to\infty}\frac{N\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i + \frac{N^N}{N!}\sum_{i=N+1}^{K} i\rho^{i-1}}{\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i + \frac{N^N}{N!}\sum_{i=N}^{K-1}\rho^i} \\ =& \frac{K}{N\mu}. \end{aligned} \tag{A.87}$$

Combining (A.85) (A.86), and (A.87), and using the definitions of k and K, we get the relation in (A.72).

To prove (A.73), recall (A.78), (A.83) and (A.84). Then,

$$\lim_{\lambda\to\infty} W_d'(\lambda) = \frac{1}{\mu^2}\lim_{\lambda\to\infty}\frac{\rho^{2k} - k^2\rho^{k+1} + (2k^2-2)\rho^k - k^2\rho^{k-1} + 1}{(\rho-1)^2(\rho^k-1)^2} = 0.$$

Furthermore, when $K = N$,

$$\lim_{\lambda\to\infty} W_p'(\lambda) = 0$$

and for $K > N$,

$$\begin{aligned}
&\lim_{\lambda\to\infty} W_p'(\lambda)\\
=&\lim_{\lambda\to\infty}\frac{\left(N\sum_{i=1}^{N-1}\frac{N^i}{i!}i\rho^{i-1} + \frac{N^N}{N!}\sum_{i=N+1}^{K} i(i-1)\rho^{i-2}\right)\left(\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i + \frac{N^N}{N!}\sum_{i=N}^{K-1}\rho^i\right)}{\left(\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i + \frac{N^N}{N!}\sum_{i=N}^{K-1}\rho^i\right)^2 N\mu^2}\\
-&\lim_{\lambda\to\infty}\frac{\left(N\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i + \frac{N^N}{N!}\sum_{i=N+1}^{K} i\rho^{i-1}\right)\left(\sum_{i=1}^{N-1}\frac{N^i}{i!}i\rho^{i-1} + \frac{N^N}{N!}\sum_{i=N}^{K-1} i\rho^{i-1}\right)}{\left(\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i + \frac{N^N}{N!}\sum_{i=N}^{K-1}\rho^i\right)^2 N\mu^2}\\
=&\frac{1}{N\mu^2}\lim_{\lambda\to\infty}\frac{\left(N\sum_{i=1}^{N-1}\frac{N^i}{i!}i\rho^{i-1} + \frac{N^N}{N!}\sum_{i=N+1}^{K} i(i-1)\rho^{i-2}\right)\left(\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i + \frac{N^N}{N!}\sum_{i=N}^{K-1}\rho^i\right)}{\left(\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i + \frac{N^N}{N!}\sum_{i=N}^{K-1}\rho^i\right)^2}\\
-&\frac{1}{N\mu^2}\lim_{\lambda\to\infty}\frac{\left(N\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i + \frac{N^N}{N!}\sum_{i=N+1}^{K} i\rho^{i-1}\right)\left(\sum_{i=1}^{N-1}\frac{N^i}{i!}i\rho^{i-1} + \frac{N^N}{N!}\sum_{i=N}^{K-1} i\rho^{i-1}\right)}{\left(\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i + \frac{N^N}{N!}\sum_{i=N}^{K-1}\rho^i\right)^2}\\
=&0.
\end{aligned}$$

This completes the proof of (A.73). □

Proof of Part (b): Note from (A.77) that $W_d(\lambda)$ can also be expressed as the following:

$$W_d(\lambda) = \frac{\sum_{i=0}^{k} i\rho^i}{\lambda\sum_{i=0}^{k-1}\rho^i} = \frac{\sum_{i=0}^{k} i\rho^i}{\mu\rho\sum_{i=0}^{k-1}\rho^i}. \tag{A.88}$$

Then, we have

$$\lim_{\lambda\to\mu} W_d(\lambda) = \frac{1}{\mu}\left(\lim_{\lambda\to\mu}\frac{\sum_{i=0}^{k} i\rho^i}{\rho\sum_{i=0}^{k-1}\rho^i}\right) = \frac{k+1}{2\mu}. \tag{A.89}$$

In addition, observe from (A.82) that when $K > N$,

$$\begin{aligned}
\lim_{\lambda\to\mu} W_p(\lambda) =& \frac{1}{N\mu}\lim_{\lambda\to\mu}\frac{N\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i + \frac{N^N}{N!}\sum_{i=N+1}^{K} i\rho^{i-1}}{\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i + \frac{N^N}{N!}\sum_{i=N}^{K-1}\rho^i} \\
=& \frac{N\sum_{i=0}^{N-1}\frac{N^i}{i!} + \frac{N^N}{N!}\sum_{i=N+1}^{K} i}{\sum_{i=0}^{N-1}\frac{N^i}{i!} + \frac{N^N}{N!}(K-N)}\left(\frac{1}{N\mu}\right) \\
=& \frac{N\sum_{i=0}^{N-1}\frac{N^i}{i!} + \frac{N^N}{N!}\frac{(K+N+1)(K-N)}{2}}{\sum_{i=0}^{N-1}\frac{N^i}{i!} + \frac{N^N}{N!}(K-N)}\left(\frac{1}{N\mu}\right) \\
=& \left(N + \frac{1}{2}\frac{\frac{N^N}{N!}(K-N+1)(K-N)}{\sum_{i=0}^{N-2}\frac{N^i}{i!} + \frac{N^N}{N!}(K-N+1)}\right)\left(\frac{1}{N\mu}\right).
\end{aligned} \tag{A.90}$$

The equation (A.90) also holds when $K = N$ since $W_p(\lambda) = \frac{1}{\mu}$ when $K = N$.

By (A.90) and the definitions of k and K from (2.5) and (2.6), we have

$$\begin{aligned}
\lim_{\lambda\to\mu} W_p(\lambda) &= \left(N + \frac{1}{2}\frac{\frac{N^N}{N!}(K-N+1)(K-N)}{\sum_{i=0}^{N-2}\frac{N^i}{i!} + \frac{N^N}{N!}(K-N+1)}\right)\frac{1}{N\mu} \\
&= \left(N + \frac{K-N}{2} - \frac{K-N}{2} + \frac{1}{2}\frac{\frac{N^N}{N!}(K-N+1)(K-N)}{\sum_{i=0}^{N-2}\frac{N^i}{i!} + \frac{N^N}{N!}(K-N+1)}\right)\frac{1}{N\mu} \\
&= \left(\frac{K+N}{2} - \frac{(K-N)\sum_{i=0}^{N-2}\frac{N^i}{i!}}{2(\sum_{i=0}^{N-2}\frac{N^i}{i!} + \frac{N^N}{N!}(K-N+1))}\right)\frac{1}{N\mu} \\
&> \left(\frac{K+N}{2} - \frac{(K-N)\sum_{i=0}^{N-2}\frac{N^i}{i!}}{2(\frac{N^N}{N!}(K-N+1))}\right)\frac{1}{N\mu} \\
&> \left(K+N - \frac{\sum_{i=0}^{N-2}\frac{N^i}{i!}}{\frac{N^N}{N!}}\right)\frac{1}{2N\mu}.
\end{aligned} \tag{A.91}$$

We already know from (A.89) that $\lim_{\lambda\to\mu} W_d(\lambda) = \frac{k+1}{2\mu}$. This and (A.91) imply that

$$\begin{aligned}
\lim_{\lambda\to\mu}[W_p(\lambda) - W_d(\lambda)] &> \left(K+N - \frac{\sum_{i=0}^{N-2}\frac{N^i}{i!}}{\frac{N^N}{N!}}\right)\frac{1}{2N\mu} - \frac{k+1}{2\mu} \\
&= \left(K - Nk - \frac{\sum_{i=0}^{N-2}\frac{N^i}{i!}}{\frac{N^N}{N!}}\right)\frac{1}{2N\mu}.
\end{aligned}$$

Therefore, if $K - Nk > \frac{\sum_{i=0}^{N-2} \frac{N^i}{i!}}{\frac{N^N}{N!}}$, $\lim_{\lambda\to\mu} [W_p(\lambda) - W_d(\lambda)] > 0$.

We now show (A.75). Suppose that $R/c > 10/\mu$. Then, $k \geq 10$ and $K > N$. Recall (A.88). Then,

$$\begin{aligned}
\lim_{\lambda\to\mu} W_d'(\lambda) =& \frac{1}{\mu^2} \lim_{\lambda\to\mu} \frac{\partial}{\partial\rho} \left(\frac{\sum_{i=0}^{k} i\rho^i}{\rho \sum_{i=0}^{k-1} \rho^i} \right) \\
=& \frac{1}{\mu^2} \lim_{\lambda\to\mu} \frac{\left(\sum_{i=1}^{k} i^2\rho^{i-1}\right)\left(\sum_{i=1}^{k} \rho^i\right) - \left(\sum_{i=0}^{k} i\rho^i\right)\left(\sum_{i=1}^{k} i\rho^{i-1}\right)}{(\sum_{i=1}^{k} \rho^i)^2} \\
=& \frac{1}{\mu^2} \frac{\frac{k(k+1)(2k+1)}{6} k - \frac{k(k+1)}{2}\frac{k(k+1)}{2}}{k^2} \\
=& \frac{k^2-1}{12\mu^2}.
\end{aligned} \tag{A.92}$$

Recalling (A.84), we now find $\lim_{\lambda\to\mu} W_p'(\lambda)$:

$$
\begin{aligned}
&\lim_{\lambda\to\mu} W_p'(\lambda) \\
&= \frac{1}{N\mu^2} \lim_{\lambda\to\mu} \frac{\left(N\sum_{i=1}^{N-1} \frac{N^i}{i!} i\rho^{i-1} + \frac{N^N}{N!}\sum_{i=N+1}^{K} i(i-1)\rho^{i-2}\right)\left(\sum_{i=0}^{N-1} \frac{N^i}{i!}\rho^i + \frac{N^N}{N!}\sum_{i=N}^{K-1}\rho^i\right)}{\left(\sum_{i=0}^{N-1} \frac{N^i}{i!}\rho^i + \frac{N^N}{N!}\sum_{i=N}^{K-1}\rho^i\right)^2} \\
&- \frac{1}{N\mu^2} \lim_{\lambda\to\mu} \frac{\left(N\sum_{i=0}^{N-1} \frac{N^i}{i!}\rho^i + \frac{N^N}{N!}\sum_{i=N+1}^{K} i\rho^{i-1}\right)\left(\sum_{i=1}^{N-1} \frac{N^i}{i!} i\rho^{i-1} + \frac{N^N}{N!}\sum_{i=N}^{K-1} i\rho^{i-1}\right)}{\left(\sum_{i=0}^{N-1} \frac{N^i}{i!}\rho^i + \frac{N^N}{N!}\sum_{i=N}^{K-1}\rho^i\right)^2} \\
&= \frac{1}{N\mu^2} \frac{\left(N\sum_{i=1}^{N-1} \frac{N^i}{i!} i + \frac{N^N}{N!}\sum_{i=N+1}^{K} i(i-1)\right)\left(\sum_{i=0}^{N-1} \frac{N^i}{i!} + \frac{N^N}{N!}(K-N)\right)}{\left(\sum_{i=0}^{N-1} \frac{N^i}{i!} + \frac{N^N}{N!}(K-N)\right)^2} \\
&- \frac{1}{N\mu^2} \frac{\left(N\sum_{i=0}^{N-1} \frac{N^i}{i!} + \frac{N^N}{N!}\sum_{i=N+1}^{K} i\right)\left(\sum_{i=1}^{N-1} \frac{N^i}{i!} i + \frac{N^N}{N!}\sum_{i=N}^{K-1} i\right)}{\left(\sum_{i=0}^{N-1} \frac{N^i}{i!} + \frac{N^N}{N!}(K-N)\right)^2} \\
&> \frac{1}{\left(\sum_{i=0}^{N-1} \frac{N^i}{i!} + \frac{N^N}{N!}(K-N)\right)^2 N\mu^2} \qquad \text{(A.93)} \\
&\cdot \left(\frac{N^N}{N!}\sum_{i=N+1}^{K} i(i-1)\right)\left(\frac{N^N}{N!}(K-N)\right) - \left(N\sum_{i=0}^{N-1}\frac{N^N}{N!} + \frac{N^N}{N!}\sum_{i=N+1}^{K} i\right)\left(\sum_{i=1}^{N-1}\frac{N^N}{N!} i + \frac{N^N}{N!}\sum_{i=N}^{K-1} i\right) \\
&= \frac{\left(\frac{N^N}{N!}\right)^2\left[\left(\sum_{i=N+1}^{K} i(i-1)(K-N)\right) - \left(\frac{K^2+N^2+K-N}{2}\right)\left(\frac{K(K-1)}{2}\right)\right]}{\left(\sum_{i=0}^{N-1} \frac{N^i}{i!} + \frac{N^N}{N!}(K-N)\right)^2 N\mu^2} \\
&= \frac{\left(\frac{N^N}{N!}\right)^2\left[\left(\frac{K(K+1)(2K+1)-N(N+1)(2N+1)}{6} - \frac{(K+N+1)(K-N)}{2}\right)(K-N) - \left(\frac{K^2+K+N^2-N}{2}\right)\left(\frac{K(K-1)}{2}\right)\right]}{\left(\sum_{i=0}^{N-1} \frac{N^i}{i!} + \frac{N^N}{N!}(K-N)\right)^2 N\mu^2} \\
&> \frac{\left(\frac{N^N}{N!}\right)^2\left[\left(\frac{K(K+1)(2K+1)-N(N+1)(2N+1)}{6} - \frac{K^2+K}{2}\right)(K-N) - \left(\frac{K^2+K+N^2-N}{2}\right)\left(\frac{K(K-1)}{2}\right)\right]}{\left(\sum_{i=0}^{N-1} \frac{N^i}{i!} + \frac{N^N}{N!}(K-N)\right)^2 N\mu^2}. \qquad \text{(A.94)}
\end{aligned}
$$

The inequality (A.93) is because $\frac{N^i}{i!} < \frac{N^N}{N!}$ for $i \in \{1, 2, \ldots, N-2\}$ and $\frac{N^{N-1}}{(N-1)!} = \frac{N^N}{N!}$. Note that the condition $R/c > 10/\mu$ in the statement of part (b) is equivalent to $k \geq 10$. Then, because $K \geq Nk$, $k \geq 10$

implies $N \leq \frac{K}{10}$. Using this and (A.94), we have

$$\begin{aligned}
&\lim_{\lambda\to\mu} W_p'(\lambda) \\
&> \frac{\left(\frac{N^N}{N!}\right)^2 \left[\left(\frac{K(K+1)(2K+1)-N(N+1)(2N+1)}{6} - \frac{K^2+K}{2}\right)(K-N) - \left(\frac{K^2+K+N^2-N}{2}\right)\left(\frac{K(K-1)}{2}\right)\right]}{\left(\sum_{i=0}^{N-1}\frac{N^i}{i!} + \frac{N^N}{N!}(K-N)\right)^2 N\mu^2} \\
&\geq \frac{\left(\frac{N^N}{N!}\right)^2 \left[\left(\frac{K(K+1)(2K+1)-\frac{K}{10}\left(\frac{K}{10}+1\right)\left(\frac{2K}{10}+1\right)}{6} - \frac{K^2+K}{2}\right)(K-N) - \left(\frac{K^2+K+(\frac{K}{10})^2-\frac{K}{10}}{2}\right)\left(\frac{K(K-1)}{2}\right)\right]}{\left(\sum_{i=0}^{N-1}\frac{N^i}{i!} + \frac{N^N}{N!}(K-N)\right)^2 N\mu^2} \qquad \text{(A.95)} \\
&\geq \frac{\left(\frac{N^N}{N!}\right)^2 \left[\left(\frac{K(K+1)(2K+1)-\frac{K}{10}\left(\frac{K}{10}+1\right)\left(\frac{2K}{10}+1\right)}{6} - \frac{K^2+K}{2}\right)(K-\frac{K}{10}) - \left(\frac{K^2+K+(\frac{K}{10})^2-\frac{K}{10}}{2}\right)\left(\frac{K(K-1)}{2}\right)\right]}{\left(\sum_{i=0}^{N-1}\frac{N^i}{i!} + \frac{N^N}{N!}(K-N)\right)^2 N\mu^2} \qquad \text{(A.96)} \\
&= \frac{\left(\frac{N^N}{N!}\right)^2 (236K^2+115K-450)K^2}{\left(\sum_{i=0}^{N-1}\frac{N^i}{i!} + \frac{N^N}{N!}(K-N)\right)^2 5000N\mu^2} \\
&> \frac{\left(\frac{N^N}{N!}\right)^2 236N^2k^2K^2}{\left(\sum_{i=0}^{N-1}\frac{N^i}{i!} + \frac{N^N}{N!}(K-N)\right)^2 5000N\mu^2} \qquad \text{(A.97)} \\
&> \frac{\left(\frac{N^N}{N!}\right)^2 K^2}{\left(\sum_{i=0}^{N-1}\frac{N^N}{N!} + \frac{N^N}{N!}(K-N)\right)^2 N\mu^2}\frac{236N^2k^2}{5000} \\
&= \frac{236Nk^2}{5000\mu^2} \qquad \text{(A.98)} \\
&> \frac{k^2-1}{12\mu^2} = \lim_{\lambda\to\mu} W_d'(\lambda). \qquad \text{(A.99)}
\end{aligned}$$

The inequality (A.95) is because $K - N > 0$, $\frac{K(K-1)}{2} > 0$, and $N \leq \frac{K}{10}$. The inequality (A.96) holds because $N \leq \frac{K}{10}$ and $\frac{K(K+1)(2K+1)-\frac{K}{10}\left(\frac{K}{10}+1\right)\left(\frac{2K}{10}+1\right)}{6} - \frac{K^2+K}{2} > 0$ as $K \geq Nk \geq 20$. The inequality (A.97) follows from the fact that $K \geq Nk$ and $115K - 450 > 0$ because $K \geq Nk \geq 20$ for $N \geq 2$ and $k \geq 10$. The inequality (A.99) follows from the fact that $N \geq 2$. Thus, $\lim_{\lambda\to\mu} W_p'(\lambda) > \lim_{\lambda\to\mu} W_d'(\lambda)$ for $k \geq 10$. Based on the inequalities above we now show (A.75). Using (A.92) and (A.98), we have

$$\lim_{\lambda\to\mu}\left[W_p'(\lambda) - W_d'(\lambda)\right] > \frac{236Nk^2}{5000\mu^2} - \frac{k^2-1}{12\mu^2} > \frac{k^2}{\mu^2}\gamma(N) > 0, \qquad \text{(A.100)}$$

where $\gamma(N) \doteq \left(\frac{236N}{5000} - \frac{1}{12}\right)$. This completes the proof of our claim in (A.75). $\square$

Proof of Part (c): Recall (A.77), (A.78), (A.79), (A.82), (A.83) and (A.84). Then, we have

$$\lim_{\lambda \to 0} W_d(\lambda) = \frac{1}{\mu} \lim_{\lambda \to 0} \frac{\left(1 - (k+1)\rho^k + k\rho^{k+1}\right)}{(\rho^k - 1)(\rho - 1)} = \frac{1}{\mu}.$$

If $k > 1$, we get

$$\lim_{\lambda \to 0} W_d'(\lambda) = \frac{1}{\mu^2} \lim_{\lambda \to 0} \frac{\rho^{2k} - k^2\rho^{k+1} + (2k^2 - 2)\rho^k - k^2\rho^{k-1} + 1}{(\rho - 1)^2(\rho^k - 1)^2} = \frac{1}{\mu^2};$$

otherwise, that is, if $k = 1$, $\lim_{\lambda \to 0} W_d'(\lambda) = 0$. In addition, for $K = N$,

$$\lim_{\lambda \to 0} W_p(\lambda) = \frac{1}{\mu} \text{ and } \lim_{\lambda \to 0} W_p'(\lambda) = 0.$$

and for $K > N$,

$$\lim_{\lambda \to 0} W_p(\lambda) = \frac{1}{N\mu} \lim_{\lambda \to 0} \frac{N \sum_{i=0}^{N-1} \frac{N^i}{i!}\rho^i + \frac{N^N}{N!} \sum_{i=N+1}^{K} i\rho^{i-1}}{\sum_{i=0}^{N-1} \frac{N^i}{i!}\rho^i + \frac{N^N}{N!} \sum_{i=N}^{K-1} \rho^i} = \frac{1}{N\mu} N = \frac{1}{\mu},$$

and

$$\begin{aligned}
&\lim_{\lambda \to 0} W_p'(\lambda) \\
&-\frac{1}{N\mu^2} \lim_{\lambda \to 0} \frac{\left(N \sum_{i=1}^{N-1} \frac{N^i}{i!} i\rho^{i-1} + \frac{N^N}{N!} \sum_{i=N+1}^{K} i(i-1)\rho^{i-2}\right)\left(\sum_{i=0}^{N-1} \frac{N^i}{i!}\rho^i + \frac{N^N}{N!} \sum_{i=N}^{K-1} \rho^i\right)}{\left(\sum_{i=0}^{N-1} \frac{N^i}{i!}\rho^i + \frac{N^N}{N!} \sum_{i=N}^{K-1} \rho^i\right)^2} \\
&-\frac{1}{N\mu^2} \lim_{\lambda \to 0} \frac{\left(N \sum_{i=0}^{N-1} \frac{N^i}{i!}\rho^i + \frac{N^N}{N!} \sum_{i=N+1}^{K} i\rho^{i-1}\right)\left(\sum_{i=1}^{N-1} \frac{N^i}{i!} i\rho^{i-1} + \frac{N^N}{N!} \sum_{i=N}^{K-1} i\rho^{i-1}\right)}{\left(\sum_{i=0}^{N-1} \frac{N^i}{i!}\rho^i + \frac{N^N}{N!} \sum_{i=N}^{K-1} \rho^i\right)^2} \qquad \text{(A.101)} \\
&= \frac{1}{N\mu^2}(N^2 - N^2) = 0. \qquad \text{(A.102)}
\end{aligned}$$

This completes the proof of part (c). $\square$

A.7 Proof of Theorem 2

Because $SW_s(R) > SW_p(R)$ by Proposition 2,

$$\frac{SW_d(R) - SW_p(R)}{SW_p(R)} > \frac{SW_d(R) - SW_s(R)}{SW_s(R)}. \tag{A.103}$$

Define

$$r \doteq \frac{R\mu}{c} - k \qquad \text{and} \qquad r_2 \doteq \frac{RN\mu}{c} - K. \tag{A.104}$$

Then $0 \le r < 1$, $0 \le r_2 < 1$ and $r_2 = rN - \lfloor rN \rfloor$. Recall from (A.8) that the social welfare in the dedicated system is

$$\begin{aligned}
SW_d(R) =& \left(\frac{1-\rho^k}{1-\rho^{k+1}}\right) RN\lambda - \left(\frac{\rho}{1-\rho} - \frac{(k+1)\rho^{k+1}}{1-\rho^{k+1}}\right) Nc \\
=& N\frac{\lambda R(\rho^k - 1)(\rho - 1) - c(\rho - (k+1)\rho^{k+1} + k\rho^{k+2})}{(\rho^{k+1}-1)(\rho-1)} \\
=& \frac{Nc}{(\rho^{k+1}-1)(\rho-1)} \left(\frac{R\mu}{c}\rho(\rho^k - 1)(\rho - 1) - (\rho - (k+1)\rho^{k+1} + k\rho^{k+2})\right) \\
=& \frac{Nc}{(\rho^{k+1}-1)(\rho-1)} \left((k+r)\rho(\rho^k - 1)(\rho - 1) - (\rho - (k+1)\rho^{k+1} + k\rho^{k+2})\right) \\
=& \frac{Nc}{(\rho^{k+1}-1)(\rho-1)} \left(r\rho^{k+2} + (1-r)\rho^{k+1} - (k+r)\rho^2 + (k+r-1)\rho\right).
\end{aligned} \tag{A.105}$$

Similarly, the social welfare in the $M/M/1/K$ system described in Lemma 7 can be expressed as follows.

$$SW_s(R) = \frac{c}{(\rho^{K+1}-1)(\rho-1)} \left(r_2\rho^{K+2} + (1-r_2)\rho^{K+1} - (K+r_2)\rho^2 + (K+r_2-1)\rho\right). \tag{A.106}$$

Define a benefit subsequence $\{R_n, n \in \mathbb{N}_+\}$ such that $R_n \doteq \frac{nc}{\mu}$ for $n \in \mathbb{N}_+$. This implies that if $R = R_n$, then $k = n$, $K = Nn$, $r = 0$ and $r_2 = 0$. Thus, (A.105) and (A.106) reduce to

$$SW_d(R_n) = \frac{Nc}{(\rho^{n+1}-1)(\rho-1)} \left(\rho^{n+1} - n\rho^2 + (n-1)\rho\right) \tag{A.107}$$

$$SW_s(R_n) = \frac{c}{(\rho^{Nn+1}-1)(\rho-1)} \left(\rho^{Nn+1} - Nn\rho^2 + (Nn-1)\rho\right). \tag{A.108}$$

These and (A.103) imply that for $\rho > 1$

$$\begin{aligned}
&\lim_{n\to\infty} \frac{SW_d(R_n) - SW_p(R_n)}{SW_p(R_n)} \\
&\geq \lim_{n\to\infty} \frac{SW_d(R_n) - SW_s(R_n)}{SW_s(R_n)} \\
&= \lim_{n\to\infty} \frac{\frac{Nc}{(\rho^{n+1}-1)(\rho-1)}\left(\rho^{n+1} - n\rho^2 + (n-1)\rho\right) - \frac{c}{(\rho^{Nn+1}-1)(\rho-1)}\left(\rho^{Nn+1} - Nn\rho^2 + (Nn-1)\rho\right)}{\frac{c}{(\rho^{Nn+1}-1)(\rho-1)}\left(\rho^{Nn+1} - Nn\rho^2 + (Nn-1)\rho\right)} \\
&= \frac{\frac{Nc}{\rho-1} - \frac{c}{\rho-1}}{\frac{c}{\rho-1}} = (N-1).
\end{aligned}$$

Thus,

$$\max_R \frac{SW_d(R) - SW_p(R)}{SW_p(R)} \geq \max_n \frac{SW_d(R_n) - SW_p(R_n)}{SW_p(R_n)} > (N-2).$$

This completes the proof. □

A.8 Proof of Theorem 3

First, let us define some constants which will be used in the remainder of the proof:

$$\eta_1 \doteq (\max\{z_0, z_4, z_5\} + 1)/\mu \quad \text{and} \quad \eta_2 \doteq (\max\{z_0, z_6, z_7, z_8\} + 1)/\mu, \tag{A.109}$$

where

$$z_0 \doteq \inf\left\{z \in \mathbb{R} : \frac{2(z+1)\rho}{\rho^{z+1}-1} < \frac{1}{4(\rho-1)^2}, \frac{z+1}{\rho^z} < \frac{1}{4} \text{ and } z > \frac{1}{\ln(\rho)} - 1\right\}, \tag{A.110}$$

$$z_4 \doteq \inf\left\{z \in \mathbb{R} : \frac{(\rho^z-1)^2}{z^2\rho^z} > 2(\rho-1)\ln(\rho) \text{ and } z > \frac{2}{\ln(\rho)}\right\}, \tag{A.111}$$

$$z_5 \doteq \inf\left\{z \in \mathbb{R} : \frac{2(z+1)}{\rho^{z-1}-1} < \frac{\rho}{2(\rho-1)^2} \text{ and } z > \frac{3}{\ln(\rho)} + 1\right\}, \tag{A.112}$$

$$z_6 \doteq \inf\left\{z \in \mathbb{R} : \frac{z}{(\rho^z-1)} < \frac{3}{8\rho^2}, z > \frac{1}{\ln(\rho)}\right\}, \tag{A.113}$$

$$z_7 \doteq \inf\left\{z \in \mathbb{R} : \frac{(\rho^z-1)}{z^3} > \ln(\rho)4\rho^2, z > \frac{3}{\ln(\rho)}\right\}, \tag{A.114}$$

$$z_8 \doteq \inf\left\{z \in \mathbb{R} : 2(z+2)\frac{1}{\rho^{(z-2)}-1} < \frac{1}{10(\rho-1)}, z > \frac{2}{\ln(\rho)} + 2\right\}. \tag{A.115}$$

We now state and prove Lemma 14 that shows the existence of constants z_0 and z_4 through z_8, which are defined in (A.110) through (A.115), respectively.

Lemma 14. *The following claims hold for $\rho > 1$:* **(a)***The constant z_4 defined in $(A.111)$ exists and it is finite.* **(b)** *The constant z_5 defined in $(A.112)$ exists and it is finite.* **(c)** *The constant z_6 defined in $(A.113)$ exists and it is finite.* **(d)** *The constant z_7 defined in $(A.114)$ exists and it is finite.* **(e)** *The constant z_8 defined in $(A.115)$ exists and it is finite.* **(f)** *The constant z_0 defined in $(A.110)$ exists and it is finite.*

Proof of Lemma 14: Part (a): Define $g_5(z) \doteq \frac{(\rho^z-1)^2}{z^2\rho^z}$. Then, note that the definition in (A.111) is equivalent to $z_4 \doteq \inf\{z \in \mathbb{R} : g_5(z) > 2(\rho-1)\ln(\rho) \text{ and } z > 2/\ln(\rho)\}$. The function $g_5(\cdot)$ is strictly increasing when $\rho > 1$ and $z > \frac{2}{\ln(\rho)}$ because

$$\begin{aligned} g_5'(z) =& \frac{2(\rho^z-1)\rho^z\ln(\rho)z^2\rho^z - (\rho^z-1)^2(2z\rho^z + z^2\rho^z\ln(\rho))}{(z^2\rho^z)^2} \\ >& \frac{2(\rho^z-1)\rho^z\ln(\rho)z^2\rho^z - \rho^z(\rho^z-1)(2z\rho^z + z^2\rho^z\ln(\rho))}{(z^2\rho^z)^2} \\ =& \frac{(\rho^z-1)z\rho^{2z}(2\ln(\rho)z - (2+\ln(\rho)z))}{(z^2\rho^z)^2} \\ =& \frac{(\rho^z-1)(\ln(\rho)z-2)}{z^3} \\ >& 0 \end{aligned} \tag{A.116}$$

for $\rho > 1$ and $z > \frac{2}{\ln(\rho)}$. In addition, when $\rho > 1$, we have

$$\lim_{z\to\infty} g_5(z) = \lim_{z\to\infty} \frac{(\rho^z-1)^2}{z^2\rho^z} = \lim_{z\to\infty} \frac{2(\rho^z-1)\rho^z\ln(\rho)}{2z\rho^z + z^2\rho^z\ln(\rho)} = \lim_{z\to\infty} \frac{2(\rho^z-1)\ln(\rho)}{2z + z^2\ln(\rho)} = \lim_{z\to\infty} \frac{2\rho^z(\ln(\rho))^2}{2 + 2z\ln(\rho)} = \infty. \tag{A.117}$$

It follows from (A.116) and (A.117) that z_4 exists and it is finite. □ **Part (b):** Define $g_6(z) \doteq \frac{2(z+1)}{\rho^{z-1}-1}$. Then, the definition in (A.112) is equivalent to $z_5 \doteq \inf\{z \in \mathbb{R} : g_6(z) < \frac{\rho}{2(\rho-1)^2} \text{ and } z > \frac{3}{\ln(\rho)} + 1\}$. The function $g_6(\cdot)$ is strictly decreasing when $\rho > 1$ and $z > \frac{3}{\ln(\rho)} + 1$ because

$$g_6'(z) = \frac{2(\rho^{z-1}-1) - 2(z+1)(\rho^{z-1})\ln(\rho)}{(\rho^{z-1}-1)^2} < \frac{2\rho^{z-1}(1-(z+1)\ln(\rho))}{(\rho^{z-1}-1)^2} < 0 \tag{A.118}$$

for $\rho > 1$ and $z > \frac{3}{\ln(\rho)} + 1$. Furthermore, when $\rho > 1$, we have

$$\lim_{z\to\infty} g_6(z) = \lim_{z\to\infty} \frac{2(z+1)}{\rho^{z-1}-1} = \lim_{z\to\infty} \frac{2}{\rho^{z-1}\ln(\rho)} = 0. \tag{A.119}$$

It follows from (A.118) and (A.119) that z_5 exists and it is finite. □ **Part (c):** Define $g_7(z) \doteq \frac{z}{(\rho^z-1)}$. Then, the definition in (A.113) is equivalent to $z_6 = \inf\{z \in \mathbb{R} : g_7(z) < \frac{3}{8\rho^2}, z > \frac{1}{\ln(\rho)}\}$. The function $g_7(\cdot)$ is strictly decreasing when $\rho > 1$ and $z > \frac{1}{\ln(\rho)}$ because

$$g_7'(z) = \frac{\rho^z - 1 - z\rho^z\ln(\rho)}{(\rho^z-1)^2} < \frac{\rho^z(1-zln(\rho))}{(\rho^z-1)^2} < 0 \tag{A.120}$$

for $\rho > 1$ and $z > \frac{1}{\ln(\rho)}$. Moreover, when $\rho > 1$, we have

$$\lim_{z\to\infty} g_7(z) = \lim_{z\to\infty} \frac{z}{(\rho^z-1)} = \lim_{z\to\infty} \frac{1}{\rho^z\ln(\rho)} = 0. \tag{A.121}$$

It follows from (A.120) and (A.121) that z_6 exists and it is finite. □ **Part (d):** Define $g_8(z) \doteq \frac{(\rho^z-1)}{z^3}$. Then, the definition in (A.114) is equivalent to $z_7 = \inf\{z \in \mathbb{R} : g_8(z) > \ln(\rho)4\rho^2, z > \frac{3}{\ln(\rho)}\}$. The function $g_8(\cdot)$ is strictly increasing when $\rho > 1$ and $z > \frac{3}{\ln(\rho)}$ because

$$g_8'(z) = \frac{\rho^z\ln(\rho)z^3 - (\rho^z-1)3z^2}{z^6} > \frac{(\ln(\rho)z-3)\rho^z}{z^4} > 0 \tag{A.122}$$

for $\rho > 1$ and $z > \frac{3}{\ln(\rho)}$. Also, for $\rho > 1$, we have

$$\lim_{z\to\infty} g_8(z) = \lim_{z\to\infty} \frac{(\rho^z-1)}{z^3} = \lim_{z\to\infty} \frac{\rho^z\ln(\rho)}{3z^2} = \lim_{z\to\infty} \frac{\rho^z(\ln(\rho))^2}{6z} = \lim_{z\to\infty} \frac{\rho^z(\ln(\rho))^3}{6} = \infty. \tag{A.123}$$

By (A.122) and (A.123), z_7 exists and it is finite. □ **Part (e):** Define $g_9(z) \doteq 2(z+2)\frac{1}{\rho^{(z-2)}-1}$. Then, the definition in (A.115) is equivalent to $z_8 = \inf\{z \in \mathbb{R} : g_9(z) < \frac{1}{10(\rho-1)}, z > \frac{2}{\ln(\rho)} + 2\}$. Note that $g_9(z)$ is strictly decreasing when $\rho > 1$ and $z > \frac{2}{\ln(\rho)} + 2$ because

$$g_9'(z) = 2\frac{(\rho^{(z-2)}-1) - (z+2)\rho^{(z-2)}\ln(\rho)}{(\rho^{(z-2)}-1)^2} < 2(\rho^{(z-2)}-1)\frac{1-(z+2)\ln(\rho)}{(\rho^{(z-2)}-1)^2} < 0 \tag{A.124}$$

for $\rho > 1$ and $z > \frac{2}{\ln(\rho)} + 2$. Furthermore, for $\rho > 1$, we have

$$\lim_{z\to\infty} g_9(z) = \lim_{z\to\infty} 2(z+2)\frac{1}{\rho^{(z-2)}-1} = \lim_{z\to\infty} \frac{2}{\rho^{(z-2)}\ln(\rho)} = 0. \tag{A.125}$$

By (A.124) and (A.125), z_8 exists and it is finite. **Part (f):** Define $g_{10}(z) \doteq \frac{2(z+1)\rho}{\rho^{z+1}-1}$ and $g_{11}(z) \doteq \frac{z+1}{\rho^z}$. Then, the definition in (A.110) is equivalent to $z_0 = \inf\{z \in \mathbb{R} : g_{10}(z) < \frac{1}{4(\rho-1)^2}, g_{11}(z) < \frac{1}{4}, z > \frac{1}{\ln(\rho)} - 1\}$. Note that both $g_{10}(z)$ and $g_{11}(z)$ are strictly decreasing when $\rho > 1$ and $z > \frac{1}{\ln(\rho)} - 1$ because

$$g'_{10}(z) = 2\rho\frac{(\rho^{z+1}-1)-(z+1)(\rho^{z+1}\ln(\rho))}{(\rho^{z+1}-1)^2} = 2\rho\frac{\rho^{z+1}(1-(z+1)\ln(\rho))-1}{(\rho^{z+1}-1)^2} < 0 \text{ and} \tag{A.126}$$

$$g'_{11}(z) = \frac{1-(z+1)\ln(\rho)}{\rho^z} < 0 \tag{A.127}$$

for $\rho > 1$ and $z > \frac{1}{\ln(\rho)} - 1$. In addition, for $\rho > 1$, we have

$$\lim_{z\to\infty} g_{10}(z) = \lim_{z\to\infty} \frac{2(z+1)\rho}{\rho^{(z+1)}-1} = \lim_{z\to\infty} \frac{2\rho}{\rho^{(z+1)}\ln(\rho)} = 0. \tag{A.128}$$

$$\lim_{z\to\infty} g_{11}(z) = \lim_{z\to\infty} \frac{z+1}{\rho^z} = \lim_{z\to\infty} \frac{1}{\rho^z\ln(\rho)} = 0. \tag{A.129}$$

By (A.126) through (A.129), z_0 exists and it is finite. □

Proof of Theorem 3 - Part (a): Take any $i \in \{1, 2, \ldots, m\}$. Recall the definition of η_1 from (A.109). First, we will show that the percentage subsequence $\beta_W(N_{i,\cdot})$ is non-negative under the stated conditions in part (a). When $N_{i,\ell} = 1$, $\frac{W_p(N_{i,\ell})-W_d(N_{i,\ell})}{W_p(N_{i,\ell})} = 0$. Recall the definition of z_1 from (A.42). Below we will show by (A.130) through (A.132) that $z_0 \geq z_1(N_{i,\ell})$ for any $N_{i,\ell} \geq 2$ when $\rho > 1$. Thus, $k > z_0$ implies $k > z_1(N_{i,\ell})$ for $\rho > 1$. Then, it follows from Theorem 1-(a)-(i) that $W_p(N_{i,\ell}) - W_d(N_{i,\ell}) > 0$ for any $N_{i,\ell} \geq 2$ and hence $\frac{W_p(N_{i,\ell})-W_d(N_{i,\ell})}{W_p(N_{i,\ell})} > 0$ when $\rho > 1$ and $k > z_0$. This and the fact that $R/c > (z_0+1)/\mu$ implies $k > z_0$ complete our argument for the statement that $\beta_W(N_{i,\cdot})$ subsequence is non-negative for any i under the stated conditions in part (a). We now show our above claim that $z_0 \geq z_1(N_{i,\ell})$ for any $N_{i,\ell} \geq 2$ when $\rho > 1$. To do so, for any $N_{i,\ell} \geq 2$, we will first show that z_0 already meets the conditions $z_1(N_{i,\ell})$

meets.

$$\frac{(z_0+1)N_{i,\ell}\rho}{\rho^{z_0+1}-1}-\frac{N_{i,\ell}-1}{4(\rho-1)^2}=(N_{i,\ell}-1)\left(\frac{N_{i,\ell}}{N_{i,\ell}-1}\frac{(z_0+1)\rho}{\rho^{z_0+1}-1}-\frac{1}{4(\rho-1)^2}\right)$$
$$\leq (N_{i,\ell}-1)\left(2\frac{(z_0+1)\rho}{\rho^{z_0+1}-1}-\frac{1}{4(\rho-1)^2}\right)$$
$$\leq 0, \tag{A.130}$$
$$\frac{(z_0+1)N_{i,\ell}}{(N_{i,\ell}-1)\rho^{z_0}}-\frac{1}{2}\leq\frac{2(z_0+1)}{\rho^{z_0}}-\frac{1}{2}\leq 0, \tag{A.131}$$
$$z_0>\frac{1}{\ln(\rho)}-1. \tag{A.132}$$

Based on these, suppose for a contradiction that $z_0 < z_1(N_{i,\ell})$ for some $N_{i,\ell} \geq 2$. Then, for any $z \in (z_0, z_1(N_{i,\ell}))$, z satisfies the set of constraints that defines $z_1(N_{i,\ell})$ since $g_1(z)$ and $g_2(z)$ (defined in the proof of Proposition 3) are strictly decreasing in z when $\rho > 1$ and $z > 1/\ln(\rho) - 1$. But, this contradicts with the definition of $z_1(N_{i,\ell})$. Thus, $z_0 \geq z_1(N_{i,\ell})$.

Next, we prove that the percentage subsequence is strictly increasing when $\rho > 1$ and $k > \max\{z_4, z_5\}$. Take any $\ell_1 \in \mathbb{N}$ and $\ell_2 \in \mathbb{N}_+$ such that $\ell_1 < \ell_2$. Let $N_1 \doteq i + \ell_1 m$ and $N_2 \doteq i + \ell_2 m$, which imply that $N_1 < N_2$ and $\{N_1, N_2\} \subset \{N_{i,\ell} : \ell = 0, 1, \ldots\}$.Based on these, the outline of the remainder of our proof is as follows. By Proposition 2, $W_p(N_2) > W_s(N_2)$. Thus,

$$W_p(N_2)-W_p(N_1)>W_s(N_2)-W_p(N_1) \tag{A.133}$$
$$=W_s(N_2)-W_s(N_1)-(W_p(N_1)-W_s(N_1)). \tag{A.134}$$

We claim and show below that if $\rho > 1$ and $k > \max\{z_4, z_5\}$,

$$W_s(N_2)-W_s(N_1)>\frac{\rho-1}{\lambda N_1(N_1+1)}\frac{\rho}{2(\rho-1)^2}, \tag{A.135}$$

and

$$W_p(N_1)-W_s(N_1)<\frac{\rho-1}{\lambda N_1(N_1+1)}\frac{\rho}{2(\rho-1)^2}. \tag{A.136}$$

Combining (A.134) through (A.136), we have

$$W_p(N_2) - W_p(N_1) > 0 \tag{A.137}$$

if $\rho > 1$ and $k > \max\{z_4, z_5\}$. Recall from (A.7) that

$$W_d(N_1) = W_d(N_2) = \frac{1}{\lambda(\rho-1)} \frac{k\rho^{k+2} - (k+1)\rho^{k+1} + \rho}{\rho^k - 1}.$$

Since $W_d(N)$ does not depend on N, by (A.137), $W_p(N_{i,\ell}) - W_d(N_{i,\ell})$ is strictly increasing in ℓ when $\rho > 1$ and $k > \max\{z_4, z_5\}$. As a result, $\beta_W(N_{i,\cdot}) = \frac{(W_p(N_{i,\cdot}) - W_d(N_{i,\cdot}))}{W_d(N_{i,\cdot})}$ is also strictly increasing in system size (i.e., ℓ) when $\rho > 1$ and $k > \max\{z_4, z_5\}$. We already know that $\frac{W_p(N_{i,\ell}) - W_d(N_{i,\ell})}{W_p(N_{i,\ell})}$ is non-negative when $\rho > 1$ and $k > z_0$. Because $R/c > \eta_1 \doteq (\max\{z_0, z_4, z_5\} + 1)/\mu$ implies $k > \max\{z_0, z_4, z_5\}$, part (a) follows.

We now show (A.135). To do so, we first derive the preliminary inequality (A.139), which will be used in later steps of the proof. Recall that $N_1 = i + \ell_1 m$ and $N_2 = i + \ell_2 m$, where $\ell_1 < \ell_2$. Then, the balking threshold in the system with N_1 servers is $K_1 = \lfloor \frac{(i+m\ell_1)R\mu}{c} \rfloor = \lfloor \frac{iR\mu}{c} \rfloor + \ell_1 \frac{Rm\mu}{c}$, whereas the corresponding figure in the one with N_2 servers is $K_2 = \lfloor \frac{(i+m\ell_2)R\mu}{c} \rfloor = \lfloor \frac{iR\mu}{c} \rfloor + \ell_2 \frac{Rm\mu}{c}$. Then,

$$\begin{aligned}
\frac{K_2}{N_2} - \frac{K_1}{N_1} = \frac{\lfloor \frac{iR\mu}{c} \rfloor + \ell_2 \frac{Rm\mu}{c}}{i + m\ell_2} - \frac{\lfloor \frac{iR\mu}{c} \rfloor + \ell_1 \frac{Rm\mu}{c}}{i + m\ell_1} &= \frac{(m\ell_1 - m\ell_2)\lfloor \frac{iR\mu}{c} \rfloor + (\ell_2 - \ell_1)\frac{iRm\mu}{c}}{(i + m\ell_1)(i + m\ell_2)} \\
&= \frac{(\ell_2 - \ell_1)(\frac{iRm\mu}{c} - m\lfloor \frac{iR\mu}{c} \rfloor)}{(i + m\ell_1)(i + m\ell_2)} \\
&\geq 0.
\end{aligned} \tag{A.138}$$

We can represent K_1 as $K_1 = N_1(k + d)$, where $0 \leq d < 1$. Thus, (A.138) implies

$$K_2 \geq \frac{N_2}{N_1} K_1 = N_2(k + d). \tag{A.139}$$

Recall from (A.11) the average sojourn time in the SQ system. Then, because $K_1 = N_1(k+d)$, in the SQ system with N_1 servers, we have

$$\begin{aligned} W_s(N_1) =& \frac{1}{N_1\lambda(\rho-1)} \frac{K_1\rho^{K_1+2} - (K_1+1)\rho^{K_1+1} + \rho}{\rho^{K_1}-1} \\ =& \frac{1}{N_1\lambda(\rho-1)} \frac{N_1(k+d)\rho^{N_1(k+d)+2} - (N_1(k+d)+1)\rho^{N_1(k+d)+1} + \rho}{\rho^{N_1(k+d)}-1} \\ =& \frac{1}{\lambda(\rho-1)} \frac{(k+d)\rho^{N_1(k+d)+2} - ((k+d)+\frac{1}{N_1})\rho^{N_1(k+d)+1} + \frac{1}{N_1}\rho}{\rho^{N_1(k+d)}-1} \\ =& \frac{1}{\lambda(\rho-1)} \left((k+d)\rho^2 - \left(k+d+\frac{1}{N_1}\right)\rho + \frac{(k+d)\rho^2 - (k+d)\rho}{\rho^{N_1(k+d)}-1} \right). \end{aligned} \tag{A.140}$$

In the SQ system with N_2 servers,

$$\begin{aligned} W_s(N_2) =& \frac{1}{N_2\lambda(\rho-1)} \frac{K_2\rho^{K_2+2} - (K_2+1)\rho^{K_2+1} + \rho}{\rho^{K_2}-1} \\ \geq& \frac{1}{N_2\lambda(\rho-1)} \frac{N_2(k+d)\rho^{N_2(k+d)+2} - (N_2(k+d)+1)\rho^{N_2(k+d)+1} + \rho}{\rho^{N_2(k+d)}-1} \qquad \text{(A.141)} \\ =& \frac{1}{\lambda(\rho-1)} \frac{(k+d)\rho^{N_2(k+d)+2} - ((k+d)+\frac{1}{N_2})\rho^{N_2(k+d)+1} + \frac{1}{N_2}\rho}{\rho^{N_2(k+d)}-1} \\ =& \frac{1}{\lambda(\rho-1)} \left((k+d)\rho^2 - \left(k+d+\frac{1}{N_2}\right)\rho + \frac{(k+d)\rho^2 - (k+d)\rho}{\rho^{N_2(k+d)}-1} \right). \qquad \text{(A.142)} \end{aligned}$$

We have (A.141) because W_s is increasing in K by (A.71) and $K_2 \geq N_2(k+d)$ from (A.139). Based on (A.140) and (A.142), we have

$$W_s(N_2) - W_s(N_1) \geq \int_{N_1}^{N_2} f_1'(\gamma) d\gamma, \tag{A.143}$$

where $f_1(\gamma) \doteq \frac{1}{\lambda(\rho-1)}\left((k+d)\rho^2-(k+d+\frac{1}{\gamma})\rho+\frac{(k+d)\rho^2-(k+d)\rho}{\rho^{\gamma(k+d)}-1}\right)$ for $\gamma \geq 1$. Note that when $k > z_4$,

$$\begin{aligned}
f_1'(\gamma) =& \frac{1}{\lambda(\rho-1)}\left(\frac{1}{\gamma^2}\rho - \frac{((k+d)\rho^2-(k+d)\rho)\rho^{\gamma(k+d)}(k+d)\ln(\rho)}{(\rho^{\gamma(k+d)}-1)^2}\right) \\
=& \frac{1}{\lambda(\rho-1)\gamma^2(\rho^{\gamma(k+d)}-1)^2}\left(\rho(\rho^{\gamma(k+d)}-1)^2-\gamma^2(k+d)^2(\rho^2-\rho)\rho^{\gamma(k+d)}\ln(\rho)\right) \\
=& \frac{\rho}{\lambda(\rho-1)\gamma^2(\rho^{\gamma(k+d)}-1)^2}\left((\rho^{\gamma(k+d)}-1)^2-\gamma^2(k+d)^2\rho^{\gamma(k+d)}(\rho-1)\ln(\rho)\right) \\
>& \frac{\rho}{\lambda(\rho-1)\gamma^2(\rho^{\gamma(k+d)}-1)^2}\frac{1}{2}(\rho^{\gamma(k+d)}-1)^2 \qquad \text{(A.144)} \\
=& \frac{\rho}{2\lambda(\rho-1)\gamma^2}. \qquad \text{(A.145)}
\end{aligned}$$

The inequality (A.144) is because when $k > z_4$ and $\gamma \geq 1$, $\gamma(k+d) > z_4$ and $\frac{(\rho^{(k+d)\gamma}-1)^2}{\gamma^2(k+d)^2\rho^{(k+d)\gamma}} > 2(\rho-1)\ln(\rho)$ by (A.111), thus

$$\begin{aligned}
&\gamma^2(k+d)^2\rho^{(k+d)\gamma}(\rho-1)\ln(\rho) - \frac{1}{2}(\rho^{(k+d)\gamma}-1)^2 \\
=& \frac{1}{2}\gamma^2(k+d)^2\rho^{(k+d)\gamma}\left(2(\rho-1)\ln(\rho) - \frac{(\rho^{(k+d)\gamma}-1)^2}{\gamma^2(k+d)^2\rho^{(k+d)\gamma}}\right) \\
<& 0.
\end{aligned}$$

It follows from (A.143) and (A.145) that if $\rho > 1$ and $k > z_4$

$$\begin{aligned}
W_s(N_2) - W_s(N_1) >& \int_{N_1}^{N_2}\frac{\rho}{2\lambda(\rho-1)\gamma^2}d\gamma = \frac{\rho}{2\lambda(\rho-1)}\left(\frac{1}{N_1}-\frac{1}{N_2}\right) \geq \frac{\rho}{2\lambda(\rho-1)}\left(\frac{1}{N_1}-\frac{1}{N_1+1}\right) \\
=& \frac{\rho-1}{\lambda N_1(N_1+1)}\frac{\rho}{2(\rho-1)^2},
\end{aligned}$$

which proves the inequality in (A.135).

We now show (A.136). Recall from (2.7) and (A.11) the average sojourn time in the pooled and SQ systems, respectively. Then, if $\rho > 1$ and $k > z_5$,

$$
\begin{aligned}
& W_p(N_1) - W_s(N_1) \\
=& \frac{\sum_{j=0}^{N_1-1} \frac{N_1^j}{j!} j\rho^j + \frac{N_1^{N_1}}{N_1!} \sum_{j=N_1}^{K_1} j\rho^j}{\left(\sum_{j=0}^{N_1-1} \frac{N_1^j}{j!}\rho^j + \frac{N_1^{N_1}}{N_1!} \sum_{j=N_1}^{K_1-1} \rho^j\right) N_1\lambda} - \frac{\sum_{j=0}^{K_1} j\rho^j}{\left(\sum_{j=0}^{K_1-1} \rho^j\right) N_1\lambda} \\
=& \frac{\sum_{j=0}^{N_1-1} \frac{N_1^j}{j!} j\rho^j + \frac{N_1^{N_1}}{N_1!} \sum_{j=N_1}^{K_1} j\rho^j}{\left(\sum_{j=0}^{N_1-1} \frac{N_1^j}{j!}\rho^j + \frac{N_1^{N_1}}{N_1!} \sum_{j=N_1}^{K_1-1} \rho^j\right) N_1\lambda} - \frac{\sum_{j=0}^{N_1-1} \frac{N_1^{N_1}}{N_1!} j\rho^j + \frac{N_1^{N_1}}{N_1!} \sum_{j=N_1}^{K_1} j\rho^j}{\left(\sum_{j=0}^{N_1-1} \frac{N_1^{N_1}}{N_1!}\rho^j + \frac{N_1^{N_1}}{N_1!} \sum_{j=N_1}^{K_1-1} \rho^j\right) N_1\lambda} \\
<& \frac{\sum_{j=0}^{N_1-1} \frac{N_1^{N_1}}{N_1!} j\rho^j + \frac{N_1^{N_1}}{N_1!} \sum_{j=N_1}^{K_1} j\rho^j}{\left(\sum_{j=0}^{N_1-1} \frac{N_1^j}{j!}\rho^j + \frac{N_1^{N_1}}{N_1!} \sum_{j=N_1}^{K_1-1} \rho^j\right) N_1\lambda} - \frac{\sum_{j=0}^{N_1-1} \frac{N_1^{N_1}}{N_1!} j\rho^j + \frac{N_1^{N_1}}{N_1!} \sum_{j=N_1}^{K_1} j\rho^j}{\left(\sum_{j=0}^{N_1-1} \frac{N_1^{N_1}}{N_1!}\rho^j + \frac{N_1^{N_1}}{N_1!} \sum_{j=N_1}^{K_1-1} \rho^j\right) N_1\lambda} \qquad \text{(A.146)} \\
=& \frac{\sum_{j=0}^{N_1-1} \frac{N_1^{N_1}}{N_1!} j\rho^j + \frac{N_1^{N_1}}{N_1!} \sum_{j=N_1}^{K_1} j\rho^j}{N_1\lambda} \\
& \cdot \frac{\sum_{j=0}^{N_1-1} (\frac{N_1^{N_1}}{N_1!} - \frac{N_1^j}{j!})\rho^j}{\left(\sum_{j=0}^{N_1-1} \frac{N_1^j}{j!}\rho^j + \frac{N_1^{N_1}}{N_1!} \sum_{j=N_1}^{K_1-1} \rho^j\right)\left(\sum_{j=0}^{N_1-1} \frac{N_1^{N_1}}{N_1!}\rho^j + \frac{N_1^{N_1}}{N_1!} \sum_{j=N_1}^{K_1-1} \rho^j\right)} \\
=& \frac{\sum_{j=0}^{K_1} j\rho^j}{\left(\sum_{j=0}^{K_1-1} \rho^j\right) N_1\lambda} \frac{\sum_{j=0}^{N_1-1} (\frac{N_1^{N_1}}{N_1!} - \frac{N_1^j}{j!})\rho^j}{\left(\sum_{j=0}^{N_1-1} \frac{N_1^j}{j!}\rho^j + \frac{N_1^{N_1}}{N_1!} \sum_{j=N_1}^{K_1-1} \rho^j\right)} \\
<& \frac{K_1}{N_1\mu} \frac{N_1(\frac{N_1^{N_1}}{N_1!})\rho^{N_1-1}}{\frac{N_1^{N_1}}{N_1!} \sum_{j=N_1}^{K_1-1} \rho^j} \qquad \text{(A.147)} \\
=& \frac{K_1}{N_1\mu} \frac{N_1\rho^{N_1-1}}{(\rho^{K_1} - \rho^{N_1})/(\rho-1)} \\
<& \frac{k+1}{\mu} \frac{N_1(\rho-1)}{\rho(\rho^{K_1-N_1} - 1)} \qquad \text{(A.148)} \\
=& \frac{k+1}{\lambda} \frac{N_1(\rho-1)}{\rho^{K_1-N_1} - 1} \\
\leq& \frac{k+1}{\lambda} \frac{N_1(\rho-1)}{\rho^{N_1k-N_1} - 1} \qquad \text{(A.149)} \\
=& \frac{\rho-1}{\lambda N_1(N_1+1)} \left(\frac{N_1^2(N_1+1)(k+1)}{\rho^{N_1k-N_1} - 1}\right) \\
\leq& \frac{\rho-1}{\lambda N_1(N_1+1)} \left(\frac{2(k+1)}{\rho^{k-1} - 1}\right) \qquad \text{(A.150)} \\
<& \frac{\rho-1}{\lambda N_1(N_1+1)} \frac{\rho}{2(\rho-1)^2}, \qquad \text{(A.151)}
\end{aligned}
$$

which proves the inequality in (A.136). Let us explain why each of the inequalities in (A.146) through (A.151) holds. The inequality (A.146) is because $\frac{N_1^j}{j!} < \frac{N_1^{N_1}}{N_1!}$ for $j \in \{1, 2, \ldots, N_1 - 2\}$ and $\frac{N_1^j}{j!} = \frac{N_1^{N_1}}{N_1!}$ for $j = N_1 - 1$. The inequality (A.147) is because

$$\frac{\sum_{j=0}^{K_1} j\rho^j}{\left(\sum_{j=0}^{K_1-1} \rho^j\right) N_1\lambda} \leq \frac{K_1 \sum_{j=1}^{K_1} \rho^j}{\left(\sum_{j=0}^{K_1-1} \rho^j\right) N_1\lambda} = \frac{K_1}{N_1\mu}$$

and for $\rho > 1$ and $j \in \{1, 2, \ldots, N_1 - 1\}$,

$$\left(\frac{N_1^{N_1}}{N_1!} - \frac{N_1^j}{j!}\right)\rho^j < \frac{N_1^{N_1}}{N_1!}\rho^{N_1-1}.$$

The inequality (A.148) follows from the fact that

$$\frac{K_1}{N_1\mu} < \frac{(k+1)N_1}{N_1\mu} = \frac{k+1}{\mu}.$$

We have (A.149) because $K_1 \geq N_1 k$ and $\rho > 1$. The inequality (A.150) is due to the fact that $f_2(\gamma) \doteq \frac{\gamma^2(\gamma+1)(k+1)}{\rho^{\gamma k-\gamma}-1}$ is decreasing in γ when $k > z_5$ and $\gamma \geq 1$, which is shown below, and hence $\frac{N_1^2(N_1+1)(k+1)}{\rho^{N_1 k-N_1}-1} \leq \frac{2(k+1)}{\rho^{k-1}-1}$. When $k > z_5$, we have $k > \frac{3}{\ln(\rho)} + 1$, and thus

$$\begin{aligned}
f_2'(\gamma) =& (k+1)\frac{(3\gamma^2+2\gamma)(\rho^{\gamma k-\gamma}-1) - \gamma^2(\gamma+1)(\rho^{\gamma k-\gamma})(k-1)\ln(\rho)}{(\rho^{\gamma k-\gamma}-1)^2} \\
<& (k+1)\frac{3\gamma^2(\gamma+1)(\rho^{\gamma k-\gamma}) - \gamma^2(\gamma+1)(\rho^{\gamma k-\gamma})(k-1)\ln(\rho)}{(\rho^{\gamma k-\gamma}-1)^2} \\
=& (k+1)\frac{\gamma^2(\gamma+1)(\rho^{\gamma k-\gamma})(3-(k-1)\ln(\rho))}{(\rho^{\gamma k-\gamma}-1)^2} \\
<& 0.
\end{aligned}$$

Finally, the inequality (A.151) is because when $\rho > 1$ and $k > z_5$, $\frac{\rho}{2(\rho-1)^2} > \frac{2(k+1)}{\rho^{k-1}-1}$ by (A.112). □

Proof of Theorem 3 - Part (b): Take any $i \in \{1, 2, \ldots, m\}$, and consider the system size subsequence $\{N_{i,\ell} = i + \ell m, \ell = 0, 1, \ldots\}$. Recall the definition of η_2 from (A.109). When $N_{i,\ell} = 1$, $\frac{SW_d(i+m\ell)-SW_p(i+m\ell)}{SW_p(i+m\ell)} = 0$. For any $N_{i,\ell} \geq 2$, according to the proof of part (a), the conditions in Theorem 1-(a) are satisfied and thus $\frac{SW_d(i+m\ell)-SW_p(i+m\ell)}{SW_p(i+m\ell)} > 0$ when $k > z_0$ and $\rho > 1$. As a result, the percentage subsequence for social welfare is non-negative under the conditions stated in part (b).

We now show that if $\rho > 1$ and $k > \max\{z_6, z_7, z_8\}$, then $\frac{SW_d(i+m\ell)-SW_p(i+m\ell)}{SW_p(i+m\ell)}$ is strictly increasing in ℓ, $\ell \in \mathbb{N}$. By the definition of the subsequence $\{N_{i,\ell}, \ell = 0, 1, \ldots\}$, the system size is equal to $i + \ell m$. Then the balking threshold in the pooled system is

$$K = d_i + \ell d_m, \tag{A.152}$$

where

$$d_i \doteq \lfloor Ri\mu/c \rfloor \quad \text{and} \quad d_m = Rm\mu/c. \tag{A.153}$$

Throughout this proof, we will include the system size as an argument of $SW_d(\cdot)$, $SW_s(\cdot)$ and $SW_p(\cdot)$; here, the index $j = s$ represents the SQ system. The following lemma identifies a bound, which we will use later in the proof.

Lemma 15. *If $\rho > 1$ and $k > \max\{z_6, z_7\}$,*

$$\frac{SW_d(i + m(\ell+1)) - SW_s(i + m(\ell+1))}{SW_s(i + m(\ell+1))} - \frac{SW_d(i+\ell m) - SW_s(i+\ell m)}{SW_s(i+\ell m)} > \frac{SW_d(i+\ell m)}{SW_s(i+\ell m)} \frac{1}{2(\ell+2)}.$$

Proof of Lemma 15: Recall from Lemmas 2 and 7 that

$$SW_d(i+\ell m) = \left(\frac{1-\rho^k}{1-\rho^{k+1}} R\lambda - \frac{\rho - (k+1)\rho^{k+1} + k\rho^{k+2}}{(\rho-1)(\rho^{k+1}-1)} c \right) (i+\ell m), \tag{A.154}$$

and

$$SW_s(i+\ell m) = \frac{1-\rho^{d_i+\ell d_m}}{1-\rho^{d_i+\ell d_m+1}} R(i+\ell m)\lambda - \frac{\rho - (d_i+\ell d_m+1)\rho^{d_i+\ell d_m+1} + (d_i+\ell d_m)\rho^{d_i+\ell d_m+2}}{(\rho-1)(\rho^{d_i+\ell d_m+1}-1)} c. \tag{A.155}$$

We will prove this lemma under each of the two possible cases about ℓ.

Case 1: First, we prove Lemma 15 for $\ell = 0$, i.e.,

$$\frac{SW_d(i+m) - SW_s(i+m)}{SW_s(i+m)} - \frac{SW_d(i) - SW_s(i)}{SW_s(i)} > \frac{SW_d(i)}{SW_s(i)} \frac{1}{4},$$

which is equivalent to

$$\frac{SW_d(i+m)}{SW_d(i)} > \frac{5}{4}\frac{SW_s(i+m)}{SW_s(i)}. \tag{A.156}$$

Using (A.154) and the fact that $1 \leq i \leq m$,

$$\frac{SW_d(i+m)}{SW_d(i)} = \frac{i+m}{i} \geq 2. \tag{A.157}$$

Recall the definition of d_i and d_m from (A.153), define

$$\bar{r}_i \doteq \frac{Ri\mu}{c} - d_i. \tag{A.158}$$

Then $0 \leq \bar{r}_i < 1$. Since $\frac{Rm\mu}{c}$ is an integer, then $\frac{R(i+\ell m)\mu}{c} - (d_i + \ell d_m) = \bar{r}_i$ for $\ell \in \mathbb{N}$.

Recall the social welfare for the scaled queuing system from (A.106). Replacing K with $d_i + \ell d_m$ and r_2 with $\bar{r}_i$ in that expression for $SW_s(N = i + \ell m)$, we have

$$\begin{aligned}
&SW_s(i+\ell m) \\
=&\left(\bar{r}_i\rho^{d_i+\ell d_m+2} + (1-\bar{r}_i)\rho^{d_i+\ell d_m+1} - (d_i+\ell d_m+\bar{r}_i)\rho^2 + (d_i+\ell d_m+\bar{r}_i-1)\rho\right) \\
&\cdot\frac{c}{(\rho^{d_i+\ell d_m+1}-1)(\rho-1)}
\end{aligned}$$

Then, for $\ell = 0$, we have the following when $\rho > 1$ and $k > z_6$:

$$\begin{aligned}
SW_s(i) =& \frac{c}{(\rho^{d_i+1}-1)(\rho-1)}\left(\bar{r}_i\rho^{d_i+2} + (1-\bar{r}_i)\rho^{d_i+1} - (d_i+\bar{r}_i)\rho^2 + (d_i+\bar{r}_i-1)\rho\right) \\
=& \frac{c}{\rho-1}\left(\bar{r}_i\rho + (1-\bar{r}_i) + \frac{\bar{r}_i\rho + (1-\bar{r}_i) - (d_i+\bar{r}_i)\rho^2 + (d_i+\bar{r}_i-1)\rho}{\rho^{d_i+1}-1}\right) \\
>& \frac{c}{\rho-1}\left(\bar{r}_i\rho + (1-\bar{r}_i) + \frac{\bar{r}_i\rho + (1-\bar{r}_i) - (d_i+1)\rho^2}{\rho^{d_i+1}-1}\right) \\
>& \frac{c}{\rho-1}\left(\frac{5}{8}(\bar{r}_i\rho + (1-\bar{r}_i)) + \frac{\bar{r}_i\rho + (1-\bar{r}_i)}{\rho^{d_i+1}-1}\right) \qquad \text{(A.159)} \\
>& \frac{5c}{8(\rho-1)}\left(\bar{r}_i\rho + (1-\bar{r}_i) + \frac{\bar{r}_i\rho + (1-\bar{r}_i)}{\rho^{d_i+1}-1}\right) \qquad \text{(A.160)}
\end{aligned}$$

The inequality (A.159) is because $\frac{(d_i+1)\rho^2}{\rho^{d_i+1}-1} < \frac{3}{8} < \frac{3}{8}(\bar{r}_i\rho + (1-\bar{r}_i))$ when $\rho > 1$ and $d_i + 1 > k > z_6$ since $\frac{z}{\rho^z-1}$ is strictly decreasing when $\rho > 1$ and $z > \frac{1}{\ln(\rho)}$ according to the proof of Lemma 14-(c) and $\frac{z_6}{\rho^{z_6}-1} < \frac{3}{8\rho^2}$ according to (A.113). Moreover,

$$\begin{aligned}
&SW_s(i+m)\\
=&\frac{c}{(\rho^{d_i+d_m+1}-1)(\rho-1)}\left(\bar{r}_i\rho^{d_i+d_m+2}+(1-\bar{r}_i)\rho^{d_i+d_m+1}-(d_i+d_m+\bar{r}_i)\rho^2+(d_i+d_m+\bar{r}_i-1)\rho\right)\\
=&\frac{c}{\rho-1}\left(\bar{r}_i\rho+(1-\bar{r}_i)+\frac{\bar{r}_i\rho+(1-\bar{r}_i)-(d_i+d_m+\bar{r}_i)\rho^2+(d_i+d_m+\bar{r}_i-1)\rho}{\rho^{d_i+d_m+1}-1}\right)\\
<&\frac{c}{\rho-1}\left(\bar{r}_i\rho+(1-\bar{r}_i)+\frac{\bar{r}_i\rho+(1-\bar{r}_i)}{\rho^{d_i+d_m+1}-1}\right)\\
<&\frac{c}{\rho-1}\left(\bar{r}_i\rho+(1-\bar{r}_i)+\frac{\bar{r}_i\rho+(1-\bar{r}_i)}{\rho^{d_i+1}-1}\right)
\end{aligned}\tag{A.161}$$

Combing (A.160) and (A.161), we have

$$\frac{SW_s(i+m)}{SW_s(i)} < \frac{8}{5}. \tag{A.162}$$

Combining (A.157) and (A.162), (A.156) follows. Thus, the lemma holds for $\ell = 0$.

Case 2: Now, we focus on the case that $\ell \geq 1$. When $\rho > 1$ and $k > \max\{z_6, z_7\}$,

$$\begin{aligned}
&SW_s(i+\ell m)\\
=&\frac{1-\rho^{d_i+\ell d_m}}{1-\rho^{d_i+\ell d_m+1}}R(i+\ell m)\lambda-\frac{\rho-(d_i+\ell d_m+1)\rho^{d_i+\ell d_m+1}+(d_i+\ell d_m)\rho^{d_i+\ell d_m+2}}{(\rho-1)(\rho^{d_i+\ell d_m+1}-1)}c\\
=&\frac{\rho^{d_i+\ell d_m}-1}{\rho^{d_i+\ell d_m+1}-1}R(i+\ell m)\lambda-\left((d_i+\ell d_m)\rho-(d_i+\ell d_m+1)+\frac{(d_i+\ell d_m+1)(\rho-1)}{\rho^{d_i+\ell d_m+1}-1}\right)\frac{c}{\rho-1}\\
=&\left(\left(\frac{1}{\rho}-\frac{1-\frac{1}{\rho}}{\rho^{d_i+\ell d_m+1}-1}\right)\frac{R(i+\ell m)\lambda}{c}-(d_i+\ell d_m)-\frac{d_i+\ell d_m+1}{\rho^{d_i+\ell d_m+1}-1}+\frac{1}{\rho-1}\right)c\\
=&\left(\left(1-\frac{\rho-1}{\rho^{d_i+\ell d_m+1}-1}\right)\frac{R(i+\ell m)\mu}{c}-(d_i+\ell d_m)-\frac{d_i+\ell d_m+1}{\rho^{d_i+\ell d_m+1}-1}+\frac{1}{\rho-1}\right)c.
\end{aligned}\tag{A.163}$$

To get a preliminary bound, we now assume that $\ell \in \mathbb{R}_+$ and $\ell \geq 1$, which imply that the social welfare is a continuous function of ℓ. Later, we will eliminate that assumption to focus on $\ell \in \mathbb{N}_+$ and use this

preliminary bound to prove the statement in Lemma 15. Taking the derivative of (A.163), we get:

$$
\begin{aligned}
&\frac{\partial SW_s(i+\ell m)}{\partial \ell} \\
&= \left(\left(1 - \frac{\rho - 1}{\rho^{d_i + \ell d_m + 1} - 1} \right) \frac{Rm\mu}{c} + \frac{\rho - 1}{(\rho^{d_i + \ell d_m + 1} - 1)^2} \rho^{d_i + \ell d_m + 1} d_m \ln(\rho) \frac{R(i+\ell m)\mu}{c} - d_m \right) c \\
&- \left(\frac{d_m(\rho^{d_i + \ell d_m + 1} - 1) - (d_i + \ell d_m + 1)\rho^{d_i + \ell d_m + 1} d_m \ln(\rho)}{(\rho^{d_i + \ell d_m + 1} - 1)^2} \right) c \\
&< \left(\left(1 - \frac{\rho - 1}{\rho^{d_i + \ell d_m + 1} - 1} \right) d_m + \frac{\rho - 1}{(\rho^{d_i + \ell d_m + 1} - 1)^2} \rho^{d_i + \ell d_m + 1} d_m \ln(\rho)(d_i + \ell d_m + 1) - d_m \right) c \\
&- \left(\frac{d_m(\rho^{d_i + \ell d_m + 1} - 1) - (d_i + \ell d_m + 1)\rho^{d_i + \ell d_m + 1} d_m \ln(\rho)}{(\rho^{d_i + \ell d_m + 1} - 1)^2} \right) c \qquad \text{(A.164)} \\
&< \left(\frac{\rho}{(\rho^{d_i + \ell d_m + 1} - 1)^2} \rho^{d_i + \ell d_m + 1} d_m \ln(\rho)(d_i + \ell d_m + 1) \right) c \\
&= \left(\frac{\rho^{d_i + \ell d_m + 2}(d_i + \ell d_m + 1) d_m \ln(\rho)}{(\rho^{d_i + \ell d_m + 1} - 1)^2} \right) c \\
&< \left(\frac{\rho^{d_i + \ell d_m + 2}(d_i + \ell d_m + 1)^2 \ln(\rho)}{(\rho^{d_i + \ell d_m + 1} - 1)^2} \right) c \qquad \text{(A.165)} \\
&\leq \left(\frac{(d_i + \ell d_m + 1)^2 \rho^2 \ln(\rho)}{(\rho - 1)(\rho^{d_i + \ell d_m + 1} - 1)} \right) c \qquad \text{(A.166)} \\
&< \frac{c}{4(\rho - 1)(d_i + \ell d_m + 1)} \qquad \text{(A.167)} \\
&\leq \frac{c}{4(\rho - 1)(\ell + 1)}. \qquad \text{(A.168)}
\end{aligned}
$$

Here, the inequality (A.164) follows from the facts that $\frac{Rm\mu}{c} = d_m$ and $\frac{R(i+\ell m)\mu}{c} < K + 1 = d_i + \ell d_m + 1$. The inequality (A.165) is because $d_m < d_i + \ell d_m + 1$ when $\ell \geq 1$. We have (A.166) because $\frac{\rho^{d_i + \ell d_m + 2}}{\rho^{d_i + \ell d_m + 1} - 1} \leq \frac{\rho^2}{\rho - 1}$. The reason for (A.167) is as follows. By the proof of Lemma 14-(d), if $\rho > 1$ and $k > z_7$, which implies $d_i + \ell d_m + 1 > K > z_7$, $\frac{\rho^{d_i + \ell d_m + 1} - 1}{(d_i + \ell d_m + 1)^3} > 4\rho^2 \ln(\rho)$, i.e., $\left(\frac{(d_i + \ell d_m + 1)^2 \rho^2 \ln(\rho)}{(\rho^{d_i + \ell d_m + 1} - 1)} \right) < \frac{1}{4(d_i + \ell d_m + 1)}$.

Based on these, we have

$$
\begin{aligned}
&\frac{\partial}{\partial \ell}\left(\frac{SW_d(i+\ell m)-SW_s(i+\ell m)}{SW_s(i+\ell m)}\right)\\
=&\frac{\partial}{\partial \ell}\left(\frac{SW_d(i+\ell m)}{SW_s(i+\ell m)}\right)\\
=&\frac{\frac{\partial SW_d(i+\ell m)}{\partial \ell}SW_s(i+\ell m)-SW_d(i+\ell m)\frac{\partial SW_s(i+\ell m)}{\partial \ell}}{(SW_s(i+\ell m))^2}\\
=&\frac{SW_d(i+\ell m)\frac{m}{i+\ell m}SW_s(i+\ell m)-SW_d(i+\ell m)\frac{\partial SW_s(i+\ell m)}{\partial \ell}}{(SW_s(i+\ell m))^2}\\
=&SW_d(i+\ell m)\frac{\frac{m}{i+\ell m}SW_s(i+\ell m)-\frac{\partial SW_s(i+\ell m)}{\partial \ell}}{(SW_s(i+\ell m))^2}\\
=&\frac{SW_d(i+\ell m)}{SW_s(i+\ell m)}\left(\frac{m}{i+\ell m}-\frac{\partial SW_s(i+\ell m)}{\partial \ell}/SW_s(i+\ell m)\right)\\
\geq&\frac{SW_d(i+\ell m)}{SW_s(i+\ell m)}\left(\frac{1}{\ell+1}-\frac{\partial SW_s(i+\ell m)}{\partial \ell}/SW_s(i+\ell m)\right) &\text{(A.169)}\\
>&\frac{SW_d(i+\ell m)}{SW_s(i+\ell m)}\left(\frac{1}{\ell+1}-\frac{c}{4(\rho-1)(\ell+1)}\Big/\frac{c}{2(\rho-1)}\right) &\text{(A.170)}\\
=&\frac{SW_d(i+\ell m)}{SW_s(i+\ell m)}\left(\frac{1}{2(\ell+1)}\right). &\text{(A.171)}
\end{aligned}
$$

The above inequality in (A.169) is because $i \leq m$, and the inequality in (A.170) follows from (A.168) and (A.173), which will be shown below. We now show (A.173). We have

$$
SW_s(i+\ell m) \geq \left(\left(1-\frac{\rho-1}{\rho^{d_i+\ell d_m+1}-1}\right)(d_i+\ell d_m)-(d_i+\ell d_m)-\frac{d_i+\ell d_m+1}{\rho^{d_i+\ell d_m+1}-1}+\frac{1}{\rho-1}\right)c \tag{A.172}
$$

$$
\begin{aligned}
&=\left(\frac{1}{\rho-1}-\frac{(\rho-1)(d_i+\ell d_m)}{\rho^{d_i+\ell d_m+1}-1}-\frac{(d_i+\ell d_m+1)}{\rho^{d_i+\ell d_m+1}-1}\right)c\\
&>\left(\frac{1}{\rho-1}-\frac{(d_i+\ell d_m+1)\rho}{\rho^{d_i+\ell d_m+1}-1}\right)c\\
&>\frac{c}{2(\rho-1)}. &\text{(A.173)}
\end{aligned}
$$

Here, the inequality (A.172) follows because (A.163) holds and $\frac{R(i+\ell m)\mu}{c} \geq d_i+\ell d_m$, and the reason for (A.173) is as follows. By the proof of Lemma 14-(c), $\frac{z}{\rho^z-1}$ is strictly decreasing when $\rho>1$ and $z>\frac{1}{\ln(\rho)}$. Also, $\frac{z_6}{\rho^{z_6}-1} \leq \frac{3}{8\rho^2} < \frac{1}{2\rho(\rho-1)}$ according to (A.113). Thus, when $\rho>1$ and $k>z_6$, $\frac{(d_i+\ell d_m+1)\rho}{\rho^{d_i+\ell d_m+1}-1} < \frac{1}{2(\rho-1)}$ since $k>z_6$ implies $d_i+\ell d_m+1>z_6$. (It is obvious that $d_i \geq k$ since $i \geq 1$.)

Based on the analysis above, we now show the statement in the lemma. If $\rho > 1$ and $k > \max\{z_6, z_7\}$,

$$\begin{aligned}
&\frac{SW_d(i+m(\ell+1)) - SW_s(i+m(\ell+1))}{SW_s(i+m(\ell+1))} - \frac{SW_d(i+\ell m) - SW_s(i+\ell m)}{SW_s(i+\ell m)} \\
>& \int_{\ell}^{\ell+1} \frac{SW_d(i+mx)}{SW_s(i+mx)} \frac{1}{2(x+1)} dx && \text{(A.174)} \\
>& \int_{\ell}^{\ell+1} \frac{SW_d(i+\ell m)}{SW_s(i+\ell m)} \frac{1}{2(\ell+2)} dx && \text{(A.175)} \\
=& \frac{SW_d(i+\ell m)}{SW_s(i+\ell m)} \frac{1}{2(\ell+2)}. && \text{(A.176)}
\end{aligned}$$

Above, (A.174) follows from (A.171); (A.175) is because $\frac{SW_d(i+mx)}{SW_s(i+mx)} \geq \frac{SW_d(i+m\ell)}{SW_s(i+m\ell)}$ as $\frac{SW_d(i+mx)}{SW_s(i+mx)}$ is increasing in x for $\rho > 1$, $k > z_7$, and $\ell \leq x \leq \ell + 1$ by (A.171). These complete the proof of Lemma 15. □

For any system size N, we have

$$
\begin{aligned}
&SW_s(N) - SW_p(N) \\
&= \left(\frac{\frac{N^N}{N!}\rho^K}{\sum_{j=0}^{N-1} \frac{N^j}{j!}\rho^j + \sum_{j=N}^{K} \frac{N^N}{N!}\rho^j} - \frac{\frac{N^N}{N!}\rho^K}{\sum_{j=0}^{N-1} \frac{N^N}{N!}\rho^j + \sum_{j=N}^{K} \frac{N^N}{N!}\rho^j} \right) RN\lambda \\
&+ \left(\frac{\sum_{j=0}^{N-1} \frac{N^j}{j!} j\rho^j + \sum_{j=N}^{K} \frac{N^N}{N!} j\rho^j}{\sum_{j=0}^{N-1} \frac{N^j}{j!}\rho^j + \sum_{j=N}^{K} \frac{N^N}{N!}\rho^j} - \frac{\sum_{j=0}^{N-1} \frac{N^N}{N!} j\rho^j + \sum_{j=N}^{K} \frac{N^N}{N!} j\rho^j}{\sum_{j=0}^{N-1} \frac{N^N}{N!}\rho^j + \sum_{j=N}^{K} \frac{N^N}{N!}\rho^j} \right) c \\
&< \frac{N^N}{N!}\rho^K \left(\frac{\sum_{j=0}^{N-1} (\frac{N^N}{N!} - \frac{N^j}{j!})\rho^j}{(\sum_{j=0}^{N-1} \frac{N^j}{j!}\rho^j + \sum_{j=N}^{K} \frac{N^N}{N!}\rho^j)(\sum_{j=0}^{N-1} \frac{N^N}{N!}\rho^j + \sum_{j=N}^{K} \frac{N^N}{N!}\rho^j)} \right) RN\lambda \\
&+ \left(\sum_{j=0}^{N-1} \frac{N^N}{N!} j\rho^j + \sum_{j=N}^{K} \frac{N^N}{N!} j\rho^j \right) \left(\frac{\sum_{j=0}^{N-1} (\frac{N^N}{N!} - \frac{N^j}{j!})\rho^j}{(\sum_{j=0}^{N-1} \frac{N^j}{j!}\rho^j + \sum_{j=N}^{K} \frac{N^N}{N!}\rho^j)(\sum_{j=0}^{N-1} \frac{N^N}{N!}\rho^j + \sum_{j=N}^{K} \frac{N^N}{N!}\rho^j)} \right) c \\
&< \left(\frac{\sum_{j=0}^{N-1} (\frac{N^N}{N!} - \frac{N^j}{j!})\rho^j}{(\sum_{j=0}^{N-1} \frac{N^j}{j!}\rho^j + \sum_{j=N}^{K} \frac{N^N}{N!}\rho^j)(\sum_{j=0}^{N-1} \frac{N^N}{N!}\rho^j + \sum_{j=N}^{K} \frac{N^N}{N!}\rho^j)} \frac{N^N}{N!}\rho^K (K+1)\rho \right) c \\
&+ \left(\left(\sum_{j=0}^{N-1} \frac{N^N}{N!} j\rho^j + \sum_{j=N}^{K} \frac{N^N}{N!} j\rho^j \right) \frac{\sum_{j=0}^{N-1} (\frac{N^N}{N!} - \frac{N^j}{j!})\rho^j}{(\sum_{j=0}^{N-1} \frac{N^j}{j!}\rho^j + \sum_{j=N}^{K} \frac{N^N}{N!}\rho^j)(\sum_{j=0}^{N-1} \frac{N^N}{N!}\rho^j + \sum_{j=N}^{K} \frac{N^N}{N!}\rho^j)} \right) c \\
&< \left(\frac{\sum_{j=0}^{N-1} \frac{N^N}{N!}\rho^j}{(\sum_{j=N}^{K} \frac{N^N}{N!}\rho^j)^2} \frac{N^N}{N!}(K+1)\rho^{K+1} \right) c \\
&+ \left(\sum_{j=0}^{N-1} \frac{N^N}{N!} j\rho^j + \sum_{j=N}^{K} \frac{N^N}{N!} j\rho^j \right) \frac{\sum_{j=0}^{N-1} \frac{N^N}{N!}\rho^j}{(\sum_{j=N}^{K} \frac{N^N}{N!}\rho^j)(\sum_{j=0}^{N-1} \frac{N^N}{N!}\rho^j + \sum_{j=N}^{K} \frac{N^N}{N!}\rho^j)} c \\
&< \left(\frac{\sum_{j=0}^{N-1} \rho^j}{(\sum_{j=N}^{K} \rho^j)^2} (K+1)\rho^{K+1} + K \frac{\sum_{j=0}^{N-1} \rho^j}{\sum_{j=N}^{K} \rho^j} \right) c \\
&= \left((K+1)\rho^{K+1} \frac{(\rho^N - 1)(\rho - 1)}{(\rho^{K+1} - \rho^N)^2} + K \frac{\rho^N - 1}{\rho^{K+1} - \rho^N} \right) c
\end{aligned} \tag{A.177}
$$

Based on this, we have the following lemma.

Lemma 16. *Recall that $N = i + \ell m$ and $K = d_i + \ell d_m$. Then, for $\rho > 1$ and $k > z_8$,*

$$SW_s(i + \ell m) - SW_p(i + \ell m) < \frac{c}{2(\rho - 1)} \frac{1}{2\ell + 5}.$$

Proof of Lemma 16: We have the the following for $\rho > 1$ and $k > z_8$:

$$
\begin{aligned}
& SW_s(i+\ell m) - SW_p(i+\ell m) \\
<& \left((d_i + \ell d_m + 1)\rho^{d_i+\ell d_m+1} \frac{(\rho^{i+\ell m}-1)(\rho-1)}{(\rho^{d_i+\ell d_m+1}-\rho^{i+\ell m})^2} + (d_i+\ell d_m)\frac{\rho^{i+\ell m}-1}{\rho^{d_i+\ell d_m+1}-\rho^{i+\ell m}} \right) c && \text{(A.178)} \\
<& \left((d_i + \ell d_m + 1)\rho^{d_i+\ell d_m+1} \frac{(\rho^{i+\ell m}-1)\rho}{(\rho^{d_i+\ell d_m+1}-\rho^{i+\ell m})^2} \right) c \\
<& \left((d_i + \ell d_m + 1)\rho^{d_i+\ell d_m+2} \frac{\rho^{i+\ell m}}{(\rho^{d_i+\ell d_m+1}-\rho^{i+\ell m})^2} \right) c \\
\leq& \left(((k+1)(i+\ell m)+1)\rho^{(k+1)(i+\ell m)+2} \frac{\rho^{i+\ell m}}{(\rho^{k(i+\ell m)+1}-\rho^{i+\ell m})^2} \right) c && \text{(A.179)} \\
=& ((k+1)(i+m\ell)+1)\frac{\rho^{(k+1)(i+m\ell)+2}}{(\rho^{k(i+m\ell)+1}-\rho^{i+m\ell})}\frac{\rho^{i+m\ell}}{(\rho^{k(i+m\ell)+1}-\rho^{i+m\ell})}c \\
=& ((k+1)(i+m\ell)+1)\rho^{i+m\ell+1}\frac{1}{1-\rho^{(i+m\ell)(1-k)-1}}\frac{1}{\rho^{(k-1)(i+m\ell)+1}-1}c \\
<& ((k+1)(i+m\ell)+1)2\rho^{i+m\ell+1}\frac{1}{\rho^{(k-1)(i+m\ell)+1}-1}c && \text{(A.180)} \\
<& 2((k+1)(i+m\ell)+1)\frac{1}{\rho^{(k-2)(i+m\ell)}-1}c \\
\leq& 2(k+2)(i+m\ell)\frac{1}{\rho^{(k-2)(i+m\ell)}-1}c && \text{(A.181)} \\
=& 2(k+2)(i+m\ell)^2\frac{1}{\rho^{(k-2)(i+m\ell)}-1}\frac{c}{i+\ell m} \\
\leq& 2(k+2)\frac{1}{\rho^{(k-2)}-1}\frac{c}{i+\ell m} && \text{(A.182)} \\
<& \frac{1}{10(\rho-1)}\frac{c}{i+\ell m} && \text{(A.183)} \\
\leq& \frac{c}{2(\rho-1)}\frac{1}{2\ell+5}. && \text{(A.184)}
\end{aligned}
$$

We now explain how we obtain the numbered inequalities above. The inequality (A.178) is due to (A.177). Because $ki \leq d_i < (k+1)i$ and $k\ell m \leq \ell d_m < (k+1)\ell m$, (A.179) follows from the fact that $d_i + \ell d_m \leq (k+1)i + (k+1)\ell m$. The inequality (A.180) is because $\rho^{(i+m\ell)(1-k)-1} \leq \rho^{(-k)} < \frac{1}{2}$ when $i + m\ell \geq 1$, $\rho > 1$ and $k > z_8 \geq \frac{2}{\ln(\rho)} + 2$. The inequality (A.181) is because $i + m\ell \geq 1$. The inequality (A.182) is due to the fact that $f_3(\gamma) \doteq \gamma^2 \frac{1}{\rho^{(k-2)\gamma-1}}$ is strictly decreasing in γ when $\rho > 1$, $\gamma \geq 1$ and $k > z_8 \geq \frac{2}{\ln(\rho)} + 2$ as shown at the end of the proof. By the proof of Lemma 14-(e), if $\rho > 1$ and $k > z_8$, $2(k+2)\frac{1}{\rho^{(k-2)}-1} < \frac{1}{10(\rho-1)}$ and thus (A.183) follows. Finally, the inequality (A.184) is because $5(i + \ell m) \geq 2\ell + 5$ when $i \in \{1, \ldots, m\}$ and $\ell = 0, 1, \ldots$.

It only remains to prove that $f_3'(\gamma) < 0$ assuming that $\gamma \in \mathbb{R}_+$ and $\gamma \geq 1$. Note that

$$\begin{aligned} f_3'(\gamma) =& \frac{2\gamma(\rho^{(k-2)\gamma}-1) - \gamma^2\rho^{(k-2)\gamma}(k-2)\ln(\rho)}{(\rho^{(k-2)\gamma-1})^2} \\ <& \gamma\rho^{(k-2)\gamma}\frac{2-\gamma(k-2)\ln(\rho)}{(\rho^{(k-2)\gamma}-1)^2} \\ \leq& \gamma\rho^{(k-2)\gamma}\frac{2-(k-2)\ln(\rho)}{(\rho^{(k-2)\gamma}-1)^2} \qquad \text{(A.185)} \\ <& 0. \end{aligned}$$

Here, (A.185) because $\gamma \geq 1$, $\rho > 1$ and $k > z_8 \geq \frac{2}{\ln(\rho)} + 2$. $\square$

We already know from (A.173) that $SW_s(i+\ell m) > \frac{c}{2(\rho-1)}$ when $\rho > 1$ and $k > \max\{z_6, z_7\}$. Combining this with Lemma 16, we have the following for $\rho > 1$ and $k > \max\{z_6, z_7, z_8\}$:

$$SW_p(i+\ell m) > SW_s(i+\ell m) - \frac{c}{2(\rho-1)}\frac{1}{2\ell+5} > \frac{c}{2(\rho-1)} - \frac{c}{2(\rho-1)}\frac{1}{2\ell+5} = \frac{c}{2(\rho-1)}\frac{2\ell+4}{2\ell+5}. \qquad \text{(A.186)}$$

This and Lemma 16 together imply that for $\rho > 1$ and $k > \max\{z_6, z_7, z_8\}$,

$$\frac{SW_s(i+\ell m)}{SW_p(i+\ell m)} = 1 + \frac{SW_s(i+\ell m) - SW_p(i+\ell m)}{SW_p(i+\ell m)} < 1 + \frac{1}{2\ell+4}. \qquad \text{(A.187)}$$

Then, if $\rho > 1$ and $k > \max\{z_6, z_7, z_8\}$, we have the following for any $i \in \{1, 2, \ldots, m\}$ and $\ell \in \mathbb{N}$:

$$\begin{aligned} & \frac{SW_d(i+m(\ell+1)) - SW_p(i+m(\ell+1))}{SW_p(i+m(\ell+1))} - \frac{SW_d(i+m\ell) - SW_p(i+m\ell)}{SW_p(i+m\ell)} \\ >& \frac{SW_d(i+m(\ell+1)) - SW_s(i+m(\ell+1))}{SW_s(i+m(\ell+1))} - \frac{SW_d(i+m\ell) - SW_p(i+m\ell)}{SW_p(i+m\ell)} \qquad \text{(A.188)} \\ =& \frac{SW_d(i+m(\ell+1)) - SW_s(i+m(\ell+1))}{SW_s(i+m(\ell+1))} - \frac{SW_d(i+m\ell) - SW_s(i+m\ell)}{SW_s(i+m\ell)} \\ +& \frac{SW_d(i+m\ell) - SW_s(i+m\ell)}{SW_s(i+m\ell)} - \frac{SW_d(i+m\ell) - SW_p(i+m\ell)}{SW_p(i+m\ell)} \\ >& \frac{SW_d(i+m\ell)}{SW_s(i+m\ell)}\left(\frac{1}{2(\ell+2)} + 1 - \frac{SW_s(i+m\ell)}{SW_p(i+m\ell)}\right) \qquad \text{(A.189)} \\ >& 0. \qquad \text{(A.190)} \end{aligned}$$

Here, the inequality (A.188) is because $SW_p(i + m(\ell + 1) < SW_s(i + m(\ell + 1))$ by Proposition 2. The inequality (A.189) is due to Lemma 15. The inequality (A.190) follows from (A.187).

Based on (A.190), for any given $i \in \mathbb{N}_+$, $\frac{SW_d(N_{i,\ell}) - SW_p(N_{i,\ell})}{SW_p(N_{i,\ell})}$ is increasing in ℓ if $\rho > 1$ and $k > \max\{z_6, z_7, z_8\}$. We already know that the subsequence is non-negative when $\rho > 1$ and $k > z_0$. This and the fact that $R/c > \eta_2 = (\max\{z_0, z_6, z_7, z_8\} + 1)/\mu$ implies $k > \max\{z_0, z_6, z_7, z_8\}$ complete the proof of the claim. □

A.9 Proof of Proposition 4

Proof of Part (a): In both dedicated and pooled systems, the service fee affects the social welfare only through balking thresholds. In the dedicated system, suppose that the fee that maximizes welfare is f_d^* and the resulting balking threshold is k_d^*. Consider another system called "dedicated help system," which is a variation of the dedicated system. In this system, there are N single-server systems. The total arrival for the system follows Poisson distribution with an arrival rate $N\lambda$, and each customer is routed to a server with probability $\frac{1}{N}$. Thus, the arrival for each single-server system is a Poisson process with rate λ. In the dedicated help system, all servers are homogeneous and service time of each server is exponentially distributed with rate μ. As soon as a single-server system has no customers, the server gets into "help" mode and he randomly chooses another single-server system in which there is at least one customer waiting (in addition to the customer the server of that queue is serving) and starts serving that customer. If there is no such system, the server stays idle until either a customer arrives to that server or a customer arrives to another queue whose server is busy. (If there is more than one server idling in the "help" mode, one of the servers, who will help, can be chosen randomly whenever a customer arrives at the queue of a busy server.) When a server starts helping another server, the help service process is interrupted and canceled altogether, and the customer being served goes back to her original queue either if a customer arrives to the queue of the server who is helping or the server who is being helped finishes his service and becomes available to serve the customer currently being served by the helper server. The new arrival to a server will be accepted if and only if the number of customers which belong to that server (including the ones that are originally routed to the queue of the server but are helped by other servers, and excluding the one (if any) that is under help of the server) n satisfies $n < k_d^*$.

Let $X_d^i(t)$ denote the number of customers that belong to the i^{th} server of the dedicated system at time t and denote by $X_h^i(t)$ the number of customers that belong to the i^{th} server of the dedicated help system at time t. (The latter excludes the customer from other servers helped by the i^{th} server, and includes the customers from the i^{th} queue helped by the other servers.) By sample path comparison, $X_d^i(t) \geq X_h^i(t)$, $t \geq 0$. Let $\lambda_{e,d}^i$ and $\lambda_{e,h}^i$ denote the throughput for the server i in the dedicated system and in the dedicated help system, respectively. Then, we have $\lambda_{e,h}^i \geq \lambda_{e,d}^i$ because if an arrival joins the i^{th} queue of the dedicated system, it implies that the number of customers in the i^{th} dedicated sub-system is less than k_d^*, and so, she will also join the help system since the number of customers customers that belong to the i^{th} server of the dedicated help system is even smaller. From this and the fact that $X_d^i(t) \geq X_h^i(t)$ $t \geq 0$, it follows that the average sojourn time for the arrivals to server i who join the dedicated help system is smaller than that who join the dedicated system, i.e., $W_h^i \leq W_d^i$.

Denote by SW_h the social welfare under the described dedicated help system. By the throughput and average sojourn time inequalities above, the dedicated help system results in larger social welfare than the dedicated system with fee f_d^*, i.e., $SW_h \geq SW_d^*$.

The dedicated help system can be thought as a kind of pooled system, but with a different admission policy. Because the socially-optimal admission control for the pooled system is a deterministic threshold policy and it can be achieved by setting a service fee, the pooled system with the socially-optimal fee (for the pooled system) will result in larger social welfare than the dedicated help system, i.e., $SW_p^* \geq SW_h$. As a result, $SW_p^* \geq SW_h \geq SW_d^*$. □

Proof of Part (b): Denote by f_d^{**} and f_p^{**} the optimal service fees that maximize the revenue in the dedicated system and the pooled system, respectively. For any fee $f \geq 0$, the revenue of the dedicated system is $RV_d(f) = \theta_d(f)f$ and the revenue of the pooled system is $RV_p(f) = \theta_p(f)f$. Replacing R with $R - f_d^{**}$ in the proof of Proposition 1 and applying the same ideas as in the proof of Proposition 1, one can show that $\theta_p(f_d^{**}) > \theta_d(f_d^{**})$. Thus, $RV_p(f_d^{**}) > RV_d(f_d^{**})$. Then, $RV_p(f_p^{**}) \geq RV_p(f_d^{**}) > RV_d(f_d^{**})$ since f_p^{**} is the fee that maximizes the revenue for the pooled system. □

A.10 Proof of Lemma 3

To prove this result, we first introduce some notation and some preliminary analysis. Let $RV_d(k)$ denote the maximum revenue that can be obtained from the dedicated system when the balking threshold of each

sub-system is k. Denote by $RV_p(K)$ the maximum revenue that can be obtained from the pooled system with the balking threshold K. At any given fee f, $k = \left\lfloor \frac{(R-f)\mu}{c} \right\rfloor$ and $K = \left\lfloor \frac{(R-f)N\mu}{c} \right\rfloor$, which imply that $f \leq R - \frac{ck}{\mu}$ when the balking threshold is k in each dedicated sub-system, and $f \leq R - \frac{cK}{N\mu}$ when the balking threshold is K in the pooled system. Using these and the throughput in both dedicated and pooled systems, we get the following expressions for $RV_d(k)$ and $RV_p(K)$. When $\rho = 1$,

$$RV_d(k) = N\lambda \left(\frac{k}{k+1} \right) \left(R - \frac{ck}{\mu} \right) \quad \text{and} \tag{A.191}$$

$$RV_p(K) = N\lambda \left(1 - \frac{\frac{N^N}{N!}}{\sum_{i=0}^{N-1} \frac{N^i}{i!} + \frac{N^N}{N!}(K+1-N)} \right) \left(R - \frac{cK}{N\mu} \right). \tag{A.192}$$

When $\rho \neq 1$,

$$RV_d(k) = N\lambda \left(\frac{1-\rho^k}{1-\rho^{k+1}} \right) \left(R - \frac{ck}{\mu} \right) \quad \text{and} \tag{A.193}$$

$$RV_p(K) = N\lambda \left(1 - \frac{\frac{N^N}{N!}\rho^K}{\sum_{i=0}^{N-1} \frac{N^i}{i!}\rho^i + \frac{N^N}{N!}\frac{\rho^{K+1}-\rho^N}{\rho-1}} \right) \left(R - \frac{cK}{N\mu} \right). \tag{A.194}$$

Lemma 17. *$RV_d(k)$ is strictly concave and thus unimodal in k; $RV_p(K)$ is strictly concave and thus unimodal in K.*

Proof of Lemma 17: When $\rho = 1$,

$$RV_d''(k) = N\lambda \left(-\frac{2}{(k+1)^3} \left(R - \frac{ck}{\mu} \right) + \frac{1}{(k+1)^2} \left(-\frac{c}{\mu} \right) + \frac{1}{(k+1)^2} \left(-\frac{c}{\mu} \right) \right) < 0$$

and

$$
\begin{aligned}
&RV_p''(K)\\
=&\left(\frac{-2(\frac{N^N}{N!})^3}{(\sum_{i=0}^{N-1}\frac{N^i}{i!}+\frac{N^N}{N!}(K+1-N))^3}\left(R-\frac{cK}{N\mu}\right)+\frac{(\frac{N^N}{N!})^2}{(\sum_{i=0}^{N-1}\frac{N^i}{i!}+\frac{N^N}{N!}(K+1-N))^2}\left(-\frac{c}{N\mu}\right)\right)N\lambda\\
+&\left(\frac{(\frac{N^N}{N!})^2}{(\sum_{i=0}^{N-1}\frac{N^i}{i!}+\frac{N^N}{N!}(K+1-N))^2}\right)\left(-\frac{c}{N\mu}\right)N\lambda\\
<&0.
\end{aligned}
$$

When $\rho \neq 1$,

$$
RV_d''(k)=(\rho-1)\ln(\rho)\frac{\rho^k\ln(\rho)(1-\rho^{2k+2})}{(1-\rho^{k+1})^4}\left(R-\frac{ck}{\mu}\right)N\lambda+2\frac{(\rho-1)\rho^k\ln(\rho)}{(1-\rho^{k+1})^2}\left(-\frac{c}{\mu}\right)N\lambda<0.
$$

Moreover, if $\rho \neq 1$, because

$$
\begin{aligned}
RV_p'(K)=&\frac{-\frac{N^N}{N!}\rho^K\ln(\rho)(\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i+\frac{N^N}{N!}\frac{-\rho^N}{\rho-1})}{(\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i+\frac{N^N}{N!}\frac{\rho^{K+1}-\rho^N}{\rho-1})^2}\left(R-\frac{cK}{N\mu}\right)N\lambda\\
&+\left(1-\frac{\frac{N^N}{N!}\rho^K}{\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i+\frac{N^N}{N!}\frac{\rho^{K+1}-\rho^N}{\rho-1}}\right)\left(-\frac{c}{N\mu}\right)N\lambda
\end{aligned}
$$

we have

$$
\begin{aligned}
&RV_p''(K)\\
=&\frac{\rho^K\ln(\rho)(\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i+\frac{N^N}{N!}\frac{\rho^{K+1}-\rho^N}{\rho-1})(\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i-\frac{N^N}{N!}\frac{\rho^{K+1}+\rho^N}{\rho-1})}{(\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i+\frac{N^N}{N!}\frac{\rho^{K+1}-\rho^N}{\rho-1})^4}\left(-\frac{N^N}{N!}\ln(\rho)\right)\\
&\left(\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i+\frac{N^N}{N!}\frac{-\rho^N}{\rho-1}\right)\left(R-\frac{cK}{N\mu}\right)N\lambda+\frac{-\frac{N^N}{N!}\rho^K\ln(\rho)(\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i+\frac{N^N}{N!}\frac{-\rho^N}{\rho-1})}{(\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i+\frac{N^N}{N!}\frac{\rho^{K+1}-\rho^N}{\rho-1})^2}\left(-\frac{c}{N\mu}\right)N\lambda\\
+&\frac{-\frac{N^N}{N!}\rho^K\ln(\rho)(\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i+\frac{N^N}{N!}\frac{-\rho^N}{\rho-1})}{(\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i+\frac{N^N}{N!}\frac{\rho^{K+1}-\rho^N}{\rho-1})^2}\left(-\frac{c}{N\mu}\right)N\lambda \qquad \text{(A.195)}\\
<&0. \qquad \text{(A.196)}
\end{aligned}
$$

The inequality (A.196) follows from the fact that each of the three terms in (A.195) are negative. The reason is that when $\rho > 1$,

$$\sum_{i=0}^{N-1} \frac{N^i}{i!}\rho^i - \frac{N^N}{N!}\frac{\rho^{K+1}+\rho^N}{\rho-1} < \frac{N^N}{N!}\frac{\rho^N-1}{\rho-1} - \frac{N^N}{N!}\frac{\rho^{K+1}+\rho^N}{\rho-1} < 0,$$
$$\sum_{i=0}^{N-1} \frac{N^i}{i!}\rho^i + \frac{N^N}{N!}\frac{-\rho^N}{\rho-1} < \frac{N^N}{N!}\frac{\rho^N-1}{\rho-1} - \frac{N^N}{N!}\frac{\rho^N}{\rho-1} < 0;$$

when $\rho < 1$,

$$\sum_{i=0}^{N-1} \frac{N^i}{i!}\rho^i - \frac{N^N}{N!}\frac{\rho^{K+1}+\rho^N}{\rho-1} > 0 \quad \text{and} \quad \sum_{i=0}^{N-1} \frac{N^i}{i!}\rho^i + \frac{N^N}{N!}\frac{-\rho^N}{\rho-1} > 0.$$

As a result, $RV_d(k)$ is strictly concave and unimodal with respect to k, and $RV_p(K)$ is strictly concave and unimodal with respect to K. □

Denote by $SW_d(k)$ and $SW_p(K)$, the social welfare under dedicated and pooled systems for any fixed balking thresholds k and K, respectively. Note that $SW_p(K)$ and $SW_d(k)$ are as in (2.8) and (A.8), and the fee affects the social welfare only through balking thresholds. We will prove the result under each of the possible two cases about ρ: $\rho = 1$ and $\rho \neq 1$. Case 1: $\rho \neq 1$. Then, $RV_d(\cdot)$ and $RV_p(\cdot)$ are as in (A.193) and (A.194), respectively, and $SW_d(\cdot)$ and $SW_p(\cdot)$ are as follows:

$$SW_d(k) = \left(\frac{1-\rho^k}{1-\rho^{k+1}}\right) RN\lambda - \left(\frac{\rho}{1-\rho} - \frac{(k+1)\rho^{k+1}}{1-\rho^{k+1}}\right) Nc,$$
$$SW_p(K) = \left(1 - \frac{\frac{N^N}{N!}\rho^K}{\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i + \frac{N^N}{N!}\sum_{i=N}^{K}\rho^i}\right) RN\lambda - \frac{\sum_{i=0}^{N-1}\frac{N^i}{i!}i\rho^i + \frac{N^N}{N!}\sum_{i=N}^{K} i\rho^i}{\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i + \frac{N^N}{N!}\sum_{i=N}^{K}\rho^i} c.$$

Note also that choosing a service fee to maximize the revenue is equivalent to choosing balking threshold (i.e., k in the dedicated system and K in the pooled system) to maximize the revenue. Define constants $\underline{r}$, $\overline{r}$, η_{11} and η_{12} as

$$\underline{r} = \eta_{11},\ \overline{r} = \eta_{12}, \tag{A.197}$$

and

$$\eta_{11} \doteq \max\left\{\left(\frac{(\rho^N-1)^2}{\rho^{N-1}(1-\rho)^2}+N-1\right)/\mu, \left(\frac{\frac{\rho-(N+1)\rho^{N+1}+N\rho^{N+2}}{(1-\rho)(1-\rho^{N+1})}N-\frac{\sum_{i=0}^{N}\frac{N^i}{i!}i\rho^i}{\sum_{i=0}^{N}\frac{N^i}{i!}\rho^i}}{\left(\frac{\frac{N^N}{N!}\rho^N}{\sum_{i=0}^{N}\frac{N^i}{i!}\rho^i}-\frac{(1-\rho)\rho^N}{1-\rho^{N+1}}\right)N\rho}\right)\Big/\mu\right\}, \tag{A.198}$$

$$\eta_{12} \doteq \min\left\{\left(\frac{(1-\rho^N)(1-\rho^{N+2})}{\rho^N(1-\rho)^2}+N+1\right)/\mu\,, \right. \tag{A.199}$$
$$\left.\left(\frac{\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i\left(\sum_{i=0}^{N}\frac{N^i}{i!}\rho^i+\frac{N^N}{N!}\rho^{N+1}\right)}{\frac{N^N}{N!}\rho^N\sum_{i=0}^{N}\frac{N^i}{i!}\rho^i-\frac{N^N}{N!}\rho^{N+1}\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i}+N+1\right)\Big/N\mu\right\}.$$

Suppose that $\frac{R}{c} \in (\eta_{11}, \eta_{12})$. Then,

$$\begin{aligned}\frac{RV_d(N)}{RV_d(N-1)} = \frac{\frac{1-\rho^N}{1-\rho^{N+1}}(R-\frac{cN}{\mu})}{\frac{1-\rho^{N-1}}{1-\rho^N}(R-\frac{c(N-1)}{\mu})} &= \left(1-\frac{\frac{c}{\mu}}{R-\frac{c(N-1)}{\mu}}\right)\frac{(1-\rho^N)^2}{(1-\rho^{N-1})(1-\rho^{N+1})}\\ &= \left(1-\frac{1}{\frac{R\mu}{c}-(N-1)}\right)\frac{(1-\rho^N)^2}{(1-\rho^{N-1})(1-\rho^{N+1})}\\ &> \left(1-\frac{\rho^{N-1}(1-\rho)^2}{(1-\rho^N)^2}\right)\frac{(1-\rho^N)^2}{(1-\rho^{N-1})(1-\rho^{N+1})} \qquad \text{(A.200)}\\ &= 1,\end{aligned}$$

which implies that

$$RV_d(N) > RV_d(N-1). \tag{A.201}$$

The inequality (A.200) above follows from the fact that $\frac{R}{c} > \eta_{11} \geq \left(\frac{(\rho^N-1)^2}{\rho^{N-1}(1-\rho)^2}+N-1\right)/\mu$. The inequality (A.201) is because $\frac{RV_d(N)}{RV_d(N-1)} > 1$ and $RV_d(N-1) = \frac{1-\rho^{N-1}}{1-\rho^N}(R-\frac{c(N-1)}{\mu}) > 0$ as $R/c > \eta_{11} > \frac{N-1}{\mu}$. Since $RV_d(N) > RV_d(N-1)$ and the revenue is unimodal in balking threshold, $RV_d(N) > RV_d(1) \geq 0$. Based on this, when $RV_d(N+1) \leq 0$, $RV_d(N) > RV_d(N+1)$. Consider the case where

$RV_d(N+1) > 0$, i.e., $R - \frac{c(N+1)}{\mu} > 0$. Then,

$$\begin{aligned}\frac{RV_d(N)}{RV_d(N+1)} = \frac{\frac{1-\rho^N}{1-\rho^{N+1}}(R-\frac{cN}{\mu})}{\frac{1-\rho^{N+1}}{1-\rho^{N+2}}(R-\frac{c(N+1)}{\mu})} &= \left(1+\frac{\frac{c}{\mu}}{R-\frac{c(N+1)}{\mu}}\right)\frac{(1-\rho^N)(1-\rho^{N+2})}{(1-\rho^{N+1})^2} \\ &= \left(1+\frac{1}{\frac{R\mu}{c}-(N+1)}\right)\frac{(1-\rho^N)(1-\rho^{N+2})}{(1-\rho^{N+1})^2} \\ &> \left(1+\frac{\rho^N(1-\rho)^2}{(1-\rho^N)(1-\rho^{N+2})}\right)\frac{(1-\rho^N)(1-\rho^{N+2})}{(1-\rho^{N+1})^2} \quad \text{(A.202)} \\ &= 1,\end{aligned}$$

which implies that

$$RV_d(N) > RV_d(N+1). \quad \text{(A.203)}$$

The inequality (A.202) holds because $\frac{R}{c} < \eta_{12} \leq \left(\frac{(1-\rho^N)(1-\rho^{N+2})}{\rho^N(1-\rho)^2} + N + 1\right)/\mu$.

As a result, since the revenue is unimodal in balking threshold by Lemma 17, by (A.201) and (A.203), $RV_d(k)$ achieves the maximum at $k = N$ and the corresponding social welfare is $SW_d(N)$ when $\frac{R}{c} \in (\eta_{11}, \eta_{12})$.

We now show a certain inequality for the revenue under the pooled system when $R/c \in (\eta_{11}, \eta_{12})$, and this inequality will be used later in the proof. There can be three cases about the sign of $RV_p(N+1)$. Case (i): Suppose that $RV_p(N+1) = \left(1 - \frac{\frac{N^N}{N!}\rho^{N+1}}{\sum_{i=0}^{N}\frac{N^i}{i!}\rho^i + \frac{N^N}{N!}\rho^{N+1}}\right)\left(R - \frac{c(N+1)}{N\mu}\right) < 0$. Then, $RV_p(N) \geq 0 > RV_p(N+1)$ since $R - \frac{c}{\mu} \geq 0$. Case (ii): Suppose that $RV_p(N+1) = \left(1 - \frac{\frac{N^N}{N!}\rho^{N+1}}{\sum_{i=0}^{N}\frac{N^i}{i!}\rho^i + \frac{N^N}{N!}\rho^{N+1}}\right)\left(R - \frac{c(N+1)}{N\mu}\right) = 0$. Then, $R - \frac{c(N+1)}{N\mu} = 0$ and $RV_p(N) = \left(1 - \frac{\frac{N^N}{N!}\rho^N}{\sum_{i=0}^{N}\frac{N^i}{i!}\rho^i}\right)\left(R - \frac{c}{\mu}\right) > 0 = RV_p(N+1)$ as $R - \frac{c}{\mu} \geq 0$. Case (iii): Suppose that $RV_p(N+1) > 0$.

Then, regarding the revenue from the pooled system, when $R/c \in (\eta_{11}, \eta_{12})$,

$$
\begin{aligned}
&\frac{RV_p(N)}{RV_p(N+1)} \\
=&\frac{\left(1-\frac{\frac{N^N}{N!}\rho^N}{\sum_{i=0}^{N}\frac{N^i}{i!}\rho^i}\right)\left(R-\frac{c}{\mu}\right)}{\left(1-\frac{\frac{N^N}{N!}\rho^{N+1}}{\sum_{i=0}^{N}\frac{N^i}{i!}\rho^i+\frac{N^N}{N!}\rho^{N+1}}\right)\left(R-\frac{c(N+1)}{N\mu}\right)} \\
=&\left(1+\frac{\frac{c}{N\mu}}{R-\frac{c(N+1)}{N\mu}}\right)\frac{1-\frac{\frac{N^N}{N!}\rho^N}{\sum_{i=0}^{N}\frac{N^i}{i!}\rho^i}}{1-\frac{\frac{N^N}{N!}\rho^{N+1}}{\sum_{i=0}^{N}\frac{N^i}{i!}\rho^i+\frac{N^N}{N!}\rho^{N+1}}} \\
=&\left(1+\frac{1}{\frac{RN\mu}{c}-(N+1)}\right)\frac{1-\frac{\frac{N^N}{N!}\rho^N}{\sum_{i=0}^{N}\frac{N^i}{i!}\rho^i}}{1-\frac{\frac{N^N}{N!}\rho^{N+1}}{\sum_{i=0}^{N}\frac{N^i}{i!}\rho^i+\frac{N^N}{N!}\rho^{N+1}}} \\
>&\left(1+\frac{\frac{N^N}{N!}\rho^N\sum_{i=0}^{N}\frac{N^i}{i!}\rho^i-\frac{N^N}{N!}\rho^{N+1}\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i}{\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i\left(\sum_{i=0}^{N}\frac{N^i}{i!}\rho^i+\frac{N^N}{N!}\rho^{N+1}\right)}\right)\left(\frac{(\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i)(\sum_{i=0}^{N}\frac{N^i}{i!}\rho^i+\frac{N^N}{N!}\rho^{N+1})}{(\sum_{i=0}^{N}\frac{N^i}{i!}\rho^i)^2}\right) \qquad \text{(A.204)} \\
=&1,
\end{aligned}
$$

which implies that

$$RV_p(N) > RV_p(N+1). \qquad \text{(A.205)}$$

((A.204) follows from $\frac{R}{c} < \eta_{12} \leq \left(\frac{\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i\left(\sum_{i=0}^{N}\frac{N^i}{i!}\rho^i+\frac{N^N}{N!}\rho^{N+1}\right)}{\frac{N^N}{N!}\rho^N\sum_{i=0}^{N}\frac{N^i}{i!}\rho^i-\frac{N^N}{N!}\rho^{N+1}\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i}+N+1\right)\Big/(N\mu)$.) Combining Cases (i) through (iii), since the revenue is unimodal in balking threshold for the pooled system as well by Lemma 17, $RV_p(K)$ achieves the maximum at $K = N$ and the corresponding social welfare is $SW_p(N)$. Then, the social welfare difference between dedicated and pooled systems under the revenue-maximizing

fee formulation satisfies the following relations when $R/c \in (\eta_{11}, \eta_{12})$:

$$
\begin{aligned}
&SW_d(N) - SW_p(N) \\
&= \left(\frac{1-\rho^N}{1-\rho^{N+1}}\right) RN\lambda - \left(\frac{\rho}{1-\rho} - \frac{(N+1)\rho^{N+1}}{1-\rho^{N+1}}\right) Nc - \left(1 - \frac{\frac{N^N}{N!}\rho^N}{\sum_{i=0}^{N} \frac{N^i}{i!}\rho^i}\right) RN\lambda + \frac{\sum_{i=0}^{N} \frac{N^i}{i!} i\rho^i}{\sum_{i=0}^{N} \frac{N^i}{i!}\rho^i} c \\
&= \left(\frac{\frac{N^N}{N!}\rho^N}{\sum_{i=0}^{N} \frac{N^i}{i!}\rho^i} - \frac{(1-\rho)\rho^N}{1-\rho^{N+1}}\right) RN\lambda - \left(\frac{\rho - (N+1)\rho^{N+1} + N\rho^{N+2}}{(1-\rho)(1-\rho^{N+1})} N - \frac{\sum_{i=0}^{N} \frac{N^i}{i!} i\rho^i}{\sum_{i=0}^{N} \frac{N^i}{i!}\rho^i}\right) c \\
&= \left(\left(\frac{\frac{N^N}{N!}\rho^N}{\sum_{i=0}^{N} \frac{N^i}{i!}\rho^i} - \frac{(1-\rho)\rho^N}{1-\rho^{N+1}}\right) \frac{R\mu}{c} N\rho - \left(\frac{\rho - (N+1)\rho^{N+1} + N\rho^{N+2}}{(1-\rho)(1-\rho^{N+1})} N - \frac{\sum_{i=0}^{N} \frac{N^i}{i!} i\rho^i}{\sum_{i=0}^{N} \frac{N^i}{i!}\rho^i}\right)\right) c \\
&> 0, \qquad \text{(A.206)}
\end{aligned}
$$

which implies that

$$SW_d(N) > SW_p(N). \qquad \text{(A.207)}$$

The inequality (A.206) follows from that fact that $\frac{R}{c} > \eta_{11} \geq \left(\frac{\frac{\rho-(N+1)\rho^{N+1}+N\rho^{N+2}}{(1-\rho)(1-\rho^{N+1})} N - \frac{\sum_{i=0}^{N} \frac{N^i}{i!} i\rho^i}{\sum_{i=0}^{N} \frac{N^i}{i!}\rho^i}}{\left(\frac{\frac{N^N}{N!}\rho^N}{\sum_{i=0}^{N} \frac{N^i}{i!}\rho^i} - \frac{(1-\rho)\rho^N}{1-\rho^{N+1}}\right) N\rho}\right) \Big/ \mu$. The inequality (A.207) completes the proof that when the fee is set to maximize the revenue, the social welfare of the dedicated system is strictly higher than that of the pooled system if $\rho \neq 1$ and $R/c \in (\eta_{11}, \eta_{12})$.

Case 2: Suppose that $\rho = 1$. Then, $RV_d(k)$ and $RV_p(K)$ are as in (A.191) and (A.192), respectively. Moreover, by (2.8) and (A.8), $SW_p(K)$ and $SW_d(k)$ are as follows:

$$
\begin{aligned}
SW_d(k) &= \left(\frac{k}{k+1}\right) RN\lambda - \left(\frac{k}{2}\right) Nc, \\
SW_p(K) &= \left(1 - \frac{\frac{N^N}{N!}\rho^K}{\sum_{i=0}^{N-1} \frac{N^i}{i!}\rho^i + \frac{N^N}{N!}\sum_{i=N}^{K} \rho^i}\right) RN\lambda - \frac{\sum_{i=0}^{N-1} \frac{N^i}{i!} i\rho^i + \frac{N^N}{N!}\sum_{i=N}^{K} i\rho^i}{\sum_{i=0}^{N-1} \frac{N^i}{i!}\rho^i + \frac{N^N}{N!}\sum_{i=N}^{K} \rho^i} c.
\end{aligned}
$$

Define constants $\underline{r}$ and $\overline{r}$ as

$$\underline{r} \doteq \max\left\{(N^2+N-1)/\mu, \left(\frac{\frac{N^2}{2} - \frac{\sum_{i=0}^{N}\frac{N^i}{i!}i}{\sum_{i=0}^{N}\frac{N^i}{i!}}}{\left(\frac{\frac{N^N}{N!}}{\sum_{i=0}^{N}\frac{N^i}{i!}} - \frac{1}{N+1}\right)N}\right)\Big/\mu\right\}, \tag{A.208}$$

$$\overline{r} \doteq \min\left\{(N^2+3N+1)/\mu, \left(\frac{\sum_{i=0}^{N-1}\frac{N^i}{i!}\left(\sum_{i=0}^{N}\frac{N^i}{i!}+\frac{N^N}{N!}\right)}{(\frac{N^N}{N!})^2} + N + 1\right)\Big/(N\mu)\right\}. \tag{A.209}$$

Using the similar ideas as in the case of $\rho \neq 1$, we can prove the following inequalities:

$$RV_d(N) > RV_d(N-1), \quad RV_d(N) > RV_d(N+1), \quad RV_p(N) > RV_p(N+1),$$
$$\text{and} \quad SW_d(N) > SW_p(N).$$

Then, because $RV_d(\cdot)$ and $RV_p(\cdot)$ are unimodal by Lemma 17, $RV_d(\cdot)$ and $RV_p(\cdot)$ both achieve their maximums at $k = N$ and $K = N$, respectively.

Combining Cases 1 and 2, $SW_d(N) > SW_p(N)$ under the revenue-maximizing fee formulation when $R/c \in (\underline{r}, \overline{r})$. It is straightforward to show that $(\underline{r}, \overline{r})$ is non-empty, for example when $N = 2$ and $\rho < \frac{\sqrt{3}-1}{2}$. □

A.11 Proof of Proposition 5

When $K > Nk$,

$$\lim_{\lambda\to\infty} W_a(\lambda) \leq \frac{k}{\mu} < \frac{K}{N\mu} = \lim_{\lambda\to\infty} W_p(\lambda)$$

The "$\leq$" follows from the fact that there are at most k customers in the system. The "$=$" above follows from (A.87). Then, the throughputs in the alternative system and in the pooled system satisfy the following:

$$\lim_{\lambda\to\infty} \theta_a(\lambda) = \lim_{\lambda\to\infty} \theta_p(\lambda) = N\mu.$$

Hence,

$$\lim_{\lambda\to\infty} SW_a(\lambda) = N\mu(R - c\lim_{\lambda\to\infty} W_a(\lambda)) > N\mu(R - c\lim_{\lambda\to\infty} W_p(\lambda)) = \lim_{\lambda\to\infty} SW_p(\lambda)$$

This completes our proof. □

A.12 Explanations and Proofs of Statements in Subsection 2.3.5

A.12.1 Preliminary Analysis

Remark A.12.1. To present the supplementary results in full generality, we will consider an unobservable system with any fixed fee $f \geq 0$. Obviously, the analysis with $f = 0$ is a special case of the analysis presented here.

Let $\widetilde{W}_j(x)$ represent the average sojourn time in the system $j \in \{d, p\}$ given that the *effective arrival rate* to a queue is x. Then, $\widetilde{W}_d(x) \doteq \infty$ if $x \geq \mu$, and $\widetilde{W}_p(x) \doteq \infty$ if $x \geq N\mu$. In line with (2.5), we consider a benefit that is not extremely small, i.e.,

$$(R - f)\mu/c > 1. \tag{A.210}$$

Specifically, the condition in (A.210) implies that the service is valuable enough that the unobservable system is not empty all the time.

Based on these, there exists a unique symmetric equilibrium such that the *equilibrium joining probability* $\widehat{q}_j$ for $j \in \{d, p\}$ is

$$\widehat{q}_j = \begin{cases} q_j^* & \text{if} \quad \widetilde{W}_j(0) < \frac{R-f}{c} < \widetilde{W}_j(\Lambda_j) \\ 1 & \text{if} \quad \widetilde{W}_j(\Lambda_j) \leq \frac{R-f}{c}, \end{cases} \tag{A.211}$$

where q_j^* is the unique solution of $R - f - c\widetilde{W}_j(\Lambda_j q_j^*) = 0$ under the stated conditions in the first line of (A.211).

Let us explain the conditions in (A.211). The condition $\widetilde{W}_j(\Lambda_j) \leq \frac{R-f}{c}$, which is equivalent to $R - f - c\widetilde{W}_j(\Lambda_j) \geq 0$, means that even if all potential customers join (i.e., the effective arrival rate is equal to the potential arrival rate), each customer gains a non-negative long-run average net benefit by joining. Thus, joining with probability 1 is the unique equilibrium strategy for all customers. We now explain the

case with $\widetilde{W}_j(0) < \frac{R-f}{c} < \widetilde{W}_j(\Lambda_j)$ in (A.211). The condition $\widetilde{W}_j(0) < \frac{R-f}{c} < \widetilde{W}_j(\Lambda_j)$ implies that if all potential customers join, each joining customer gets a negative long-run average net benefit. Because a customer is better off by balking in that case, joining with probability 1 cannot be an equilibrium strategy. The aforementioned condition also implies that if none of the potential customers join, a customer is better off by joining the queue as its long-run average net benefit would be non-negative in that case. Thus, balking with probability 1 cannot be an equilibrium strategy either. The unique equilibrium strategy is such that the joining probability is the solution of $R - f - c\widetilde{W}_j(\Lambda_j q_j^*) = 0$, which makes the long-run average net benefit zero. Note that (A.211) does not include the case $\widetilde{W}_j(0) \geq \frac{R-f}{c}$. This is because (A.210) implies that $\widetilde{W}_j(0) < \frac{R-f}{c}$.

Based on (A.211), in equilibrium, the effective arrival rate to a queue is

$$\widehat{\lambda}_{e,j} = \Lambda_j \widehat{q}_j, \quad j \in \{d, p\}, \tag{A.212}$$

and the average sojourn time in the system j is

$$\widehat{W}_j = \widetilde{W}_j(\widehat{\lambda}_{e,j}), \quad j \in \{d, p\}. \tag{A.213}$$

It is worth noting that $\lambda \widehat{q}_j < \mu$, and thus, in equilibrium, each system $j \in \{d, p\}$ is stable regardless of the fact that $\rho < 1$ or $\rho \geq 1$.

We first state and prove Lemmas 18 and 19 that will be used in proving results in Section 2.3.5. For these lemmas, recall the notation $\rho \doteq \lambda/\mu$.

Lemma 18. *Recall Remark A.12.1. In the unobservable pooled system, the average sojourn time and social welfare in equilibrium are respectively given by*

$$\widehat{W}_p = \begin{cases} \frac{R-f}{c} & \text{if } \rho < 1 \text{ and } \frac{R-f}{c} < \frac{\sum_{i=0}^{N-1} \frac{N^i}{i!} i\rho^i + \frac{N^N}{N!} \sum_{i=N}^{\infty} i\rho^i}{\left(\sum_{i=0}^{N-1} \frac{N^i}{i!} \rho^i + \frac{N^N}{N!} \sum_{i=N}^{\infty} \rho^i\right) N\lambda}, \text{ or } \rho \geq 1, \\ \frac{\sum_{i=0}^{N-1} \frac{N^i}{i!} i\rho^i + \frac{N^N}{N!} \sum_{i=N}^{\infty} i\rho^i}{\left(\sum_{i=0}^{N-1} \frac{N^i}{i!} \rho^i + \frac{N^N}{N!} \sum_{i=N}^{\infty} \rho^i\right) N\lambda} & \text{if } \rho < 1 \text{ and } \frac{\sum_{i=0}^{N-1} \frac{N^i}{i!} i\rho^i + \frac{N^N}{N!} \sum_{i=N}^{\infty} i\rho^i}{\left(\sum_{i=0}^{N-1} \frac{N^i}{i!} \rho^i + \frac{N^N}{N!} \sum_{i=N}^{\infty} \rho^i\right) N\lambda} \leq \frac{R-f}{c}, \end{cases} \tag{A.214}$$

$$\widehat{SW}_p = \begin{cases} N\lambda q_p^* f, & \text{if } \rho < 1 \text{ and } \frac{R-f}{c} < \frac{\sum_{i=0}^{N-1} \frac{N^i}{i!} i\rho^i + \frac{N^N}{N!} \sum_{i=N}^{\infty} i\rho^i}{\left(\sum_{i=0}^{N-1} \frac{N^i}{i!} \rho^i + \frac{N^N}{N!} \sum_{i=N}^{\infty} \rho^i\right) N\lambda}, \text{ or } \rho \geq 1 \\ RN\lambda - c\frac{\sum_{i=0}^{N-1} \frac{N^i}{i!} i\rho^i + \frac{N^N}{N!} \sum_{i=N}^{\infty} i\rho^i}{\sum_{i=0}^{N-1} \frac{N^i}{i!} \rho^i + \frac{N^N}{N!} \sum_{i=N}^{\infty} \rho^i}, & \text{if } \rho < 1 \text{ and } \frac{\sum_{i=0}^{N-1} \frac{N^i}{i!} i\rho^i + \frac{N^N}{N!} \sum_{i=N}^{\infty} i\rho^i}{\left(\sum_{i=0}^{N-1} \frac{N^i}{i!} \rho^i + \frac{N^N}{N!} \sum_{i=N}^{\infty} \rho^i\right) N\lambda} \leq \frac{R-f}{c} \end{cases} \tag{A.215}$$

where q_p^* is the equilibrium joining probability and satisfies the following equation under the stated conditions in the first line of $(A.215)$*:*

$$\widetilde{W}_p(N\lambda q_p^*) = \frac{\sum_{i=0}^{N-1} \frac{N^i}{i!} i(\lambda q_p^*/\mu)^i + \frac{N^N}{N!} \sum_{i=N}^{\infty} i(\lambda q_p^*/\mu)^i}{\left(\sum_{i=0}^{N-1} \frac{N^i}{i!} (\lambda q_p^*/\mu)^i + \frac{N^N}{N!} \sum_{i=N}^{\infty} (\lambda q_p^*/\mu)^i\right) N\lambda q_p^*} = \frac{R-f}{c}. \tag{A.216}$$

Remark A.12.2. By the proof of Lemma 18, there exists a unique q_p^* that satisfies (A.216) if $\rho \geq 1$, or $\rho < 1$ and $\frac{R-f}{c} < \left(\sum_{i=0}^{N-1} \frac{N^i}{i!} i\rho^i + \frac{N^N}{N!} \sum_{i=N}^{\infty} i\rho^i\right) / \left(\left(\sum_{i=0}^{N-1} \frac{N^i}{i!} \rho^i + \frac{N^N}{N!} \sum_{i=N}^{\infty} \rho^i\right) N\lambda\right)$.

Proof of Lemma 18: Consider the unobservable pooled system, which is an $M/M/N$ queueing system. Recall that the service rate of each server is μ and suppose that the effective arrival rate is $x < N\mu$. Then, the stationary probability distribution of the number of customers in this system is as follows (see Section 7.3.3 of Kulkarni (2010)):

$$\hat{\pi}_0(x) = \left(\sum_{i=0}^{N-1} \frac{N^i}{i!} \rho_x^i + \frac{N^N}{N!} \sum_{i=N}^{\infty} \rho_x^i\right)^{-1}, \tag{A.217}$$

$$\hat{\pi}_i(x) = \hat{\pi}_0(x) N^i \rho_x^i / i! \quad \text{for } i = 1, \ldots, N \quad \text{and} \quad \hat{\pi}_i(x) = \hat{\pi}_0(x) N^N \rho_x^i / N! \quad \text{for } i = N+1, N+2 \ldots \tag{A.218}$$

where $\rho_x \doteq \frac{x}{N\mu}$. Then, in this system, the long-run average number of customers is

$$\widetilde{L}_p(x) = \sum_{i=0}^{\infty} \hat{\pi}_i(x) i = \frac{\sum_{i=0}^{N-1} \frac{N^i}{i!} i\rho_x^i + \frac{N^N}{N!} \sum_{i=N}^{\infty} i\rho_x^i}{\sum_{i=0}^{N-1} \frac{N^i}{i!} \rho_x^i + \frac{N^N}{N!} \sum_{i=N}^{\infty} \rho_x^i}.$$

Because $\widetilde{W}_p(x) = \widetilde{L}_p(x)/x$ by Little's law, we have

$$\widetilde{W}_p(x) = \sum_{i=0}^{\infty} \hat{\pi}_i(x) i/x = \frac{\sum_{i=0}^{N-1} \frac{N^i}{i!} i\rho_x^i + \frac{N^N}{N!} \sum_{i=N}^{\infty} i\rho_x^i}{(\sum_{i=0}^{N-1} \frac{N^i}{i!} \rho_x^i + \frac{N^N}{N!} \sum_{i=N}^{\infty} \rho_x^i) x}.$$

Based on this, if the effective arrival rate is Λ_p and $\Lambda_p = N\lambda < N\mu$, which is equivalent to $\rho < 1$,

$$\widetilde{W}_p(\Lambda_p) = \widetilde{W}_p(N\lambda) = \frac{\sum_{i=0}^{N-1} \frac{N^i}{i!} i\rho^i + \frac{N^N}{N!} \sum_{i=N}^{\infty} i\rho^i}{(\sum_{i=0}^{N-1} \frac{N^i}{i!} \rho^i + \frac{N^N}{N!} \sum_{i=N}^{\infty} \rho^i) N\lambda}.$$

On the other hand, if the effective arrival rate is Λ_p and $\rho \geq 1$, we have $\widetilde{W}_p(\Lambda_p) = \infty$. Combining these two cases with the fact that $\widetilde{W}_p(0) = 1/\mu$, it follows from (A.211) and (A.212) that the effective arrival rate *in equilibrium* is

$$\widehat{\lambda}_{e,p} = \begin{cases} N\lambda q_p^* & \text{if } \rho < 1 \text{ and } \frac{R-f}{c} < \frac{\sum_{i=0}^{N-1} \frac{N^i}{i!} i\rho^i + \frac{N^N}{N!} \sum_{i=N}^{\infty} i\rho^i}{(\sum_{i=0}^{N-1} \frac{N^i}{i!} \rho^i + \frac{N^N}{N!} \sum_{i=N}^{\infty} \rho^i) N\lambda}, \text{ or } \rho \geq 1, \\ N\lambda & \text{if } \rho < 1 \text{ and } \frac{\sum_{i=0}^{N-1} \frac{N^i}{i!} i\rho^i + \frac{N^N}{N!} \sum_{i=N}^{\infty} i\rho^i}{(\sum_{i=0}^{N-1} \frac{N^i}{i!} \rho^i + \frac{N^N}{N!} \sum_{i=N}^{\infty} \rho^i) N\lambda} \leq \frac{R-f}{c}, \end{cases} \tag{A.219}$$

where the equilibrium joining probability q_p^* is chosen such that $\widetilde{W}_p(N\lambda q_p^*) = \frac{R-f}{c}$. (It is perhaps worth noting that we did not include the condition $1/\mu < (R-f)/c$ in the first line of (A.219) because (A.210) already implies that.) The solution q_p^* exists and is unique if $\widetilde{W}_p(N\lambda) > (R-f)/c$, which is equivalent to the conditions in the first line of (A.219). The reason is as follows. It is shown at the end of this proof that $\widetilde{W}_p(Ny)$ strictly increases with y for $y < \mu$. We already know that $\widetilde{W}_p(\cdot)$ is a continuous function for $y < \mu$. These and the facts that $\lim_{y\to 0} \widetilde{W}_p(y) = 1/\mu < (R-f)/c$ by (A.210) and $\widetilde{W}_p(N\lambda) > (R-f)/c$ imply the existence and the uniqueness of q_p^*.

Based on these, by (A.213), in equilibrium, the long-run average sojourn time is

$$\widehat{W}_p = \begin{cases} \frac{R-f}{c} & \text{if } \rho < 1 \text{ and } \frac{R-f}{c} < \frac{\sum_{i=0}^{N-1} \frac{N^i}{i!} i\rho^i + \frac{N^N}{N!} \sum_{i=N}^{\infty} i\rho^i}{(\sum_{i=0}^{N-1} \frac{N^i}{i!} \rho^i + \frac{N^N}{N!} \sum_{i=N}^{\infty} \rho^i) N\lambda}, \text{ or } \rho \geq 1, \\ \frac{\sum_{i=0}^{N-1} \frac{N^i}{i!} i\rho^i + \frac{N^N}{N!} \sum_{i=N}^{\infty} i\rho^i}{(\sum_{i=0}^{N-1} \frac{N^i}{i!} \rho^i + \frac{N^N}{N!} \sum_{i=N}^{\infty} \rho^i) N\lambda} & \text{if } \rho < 1 \text{ and } \frac{\sum_{i=0}^{N-1} \frac{N^i}{i!} i\rho^i + \frac{N^N}{N!} \sum_{i=N}^{\infty} i\rho^i}{(\sum_{i=0}^{N-1} \frac{N^i}{i!} \rho^i + \frac{N^N}{N!} \sum_{i=N}^{\infty} \rho^i) N\lambda} \leq \frac{R-f}{c}. \end{cases}$$

By Little's Law, $\widehat{L}_p = \widehat{\lambda}_{e,p} \widehat{W}_p$. Therefore, in equilibrium, the long-run average number of customers in the system is

$$\widehat{L}_p = \begin{cases} N\lambda q_p^* \frac{R-f}{c} & \text{if } \rho < 1 \text{ and } \frac{R-f}{c} < \frac{\sum_{i=0}^{N-1} \frac{N^i}{i!} i\rho^i + \frac{N^N}{N!} \sum_{i=N}^{\infty} i\rho^i}{(\sum_{i=0}^{N-1} \frac{N^i}{i!} \rho^i + \frac{N^N}{N!} \sum_{i=N}^{\infty} \rho^i) N\lambda}, \text{ or } \rho \geq 1, \\ \frac{\sum_{i=0}^{N-1} \frac{N^i}{i!} i\rho^i + \frac{N^N}{N!} \sum_{i=N}^{\infty} i\rho^i}{\sum_{i=0}^{N-1} \frac{N^i}{i!} \rho^i + \frac{N^N}{N!} \sum_{i=N}^{\infty} \rho^i} & \text{if } \rho < 1 \text{ and } \frac{\sum_{i=0}^{N-1} \frac{N^i}{i!} i\rho^i + \frac{N^N}{N!} \sum_{i=N}^{\infty} i\rho^i}{(\sum_{i=0}^{N-1} \frac{N^i}{i!} \rho^i + \frac{N^N}{N!} \sum_{i=N}^{\infty} \rho^i) N\lambda} \leq \frac{R-f}{c}. \end{cases}$$

Recall the social welfare from (2.19). Then, using the expressions above, (A.215) immediately follow.

We now show that $\widetilde{W}_p(Ny)$ is strictly increasing with y. Note that

$$\begin{aligned}
\widetilde{W}_p(Ny) =& \frac{\sum_{i=0}^{N-1} \frac{N^i}{i!} i(y/\mu)^i + \frac{N^N}{N!} \sum_{i=N}^{\infty} i(y/\mu)^i}{\left(\sum_{i=0}^{N-1} \frac{N^i}{i!}(y/\mu)^i + \frac{N^N}{N!} \sum_{i=N}^{\infty}(y/\mu)^i\right) Ny} \\
=& \frac{1}{N\mu} \frac{\sum_{i=1}^{N-1} \frac{N^i}{(i-1)!}(y/\mu)^{i-1} + \frac{N^N}{N!} \sum_{i=N}^{\infty} i(y/\mu)^{i-1}}{\sum_{i=0}^{N-1} \frac{N^i}{i!}(y/\mu)^i + \frac{N^N}{N!} \sum_{i=N}^{\infty}(y/\mu)^i} \\
=& \frac{1}{N\mu} \frac{N\sum_{i=0}^{N-2} \frac{N^i}{i!}(y/\mu)^i + \frac{N^N}{N!} \sum_{i=N-1}^{\infty}(i+1)(y/\mu)^i}{\sum_{i=0}^{N-1} \frac{N^i}{i!}(y/\mu)^i + \frac{N^N}{N!} \sum_{i=N}^{\infty}(y/\mu)^i} \\
=& \frac{1}{N\mu}\left(N + \frac{\frac{N^N}{N!} \sum_{i=N}^{\infty}(i+1-N)(y/\mu)^i}{\sum_{i=0}^{N-1} \frac{N^i}{i!}(y/\mu)^i + \frac{N^N}{N!} \sum_{i=N}^{\infty}(y/\mu)^i}\right) \\
=& \frac{1}{N\mu}\left(N + \frac{\frac{N^N}{N!}\frac{(y/\mu)^N}{(1-y/\mu)^2}}{\sum_{i=0}^{N-1} \frac{N^i}{i!}(y/\mu)^i + \frac{N^N}{N!}\frac{(y/\mu)^N}{1-y/\mu}}\right) \\
=& \frac{1}{N\mu}\left(N + \frac{\frac{N^N}{N!}}{\sum_{i=0}^{N-1} \frac{N^i}{i!}(y/\mu)^{i-N}(1-y/\mu)^2 + \frac{N^N}{N!}(1-y/\mu)}\right).
\end{aligned}$$

Because $\sum_{i=0}^{N-1} \frac{N^i}{i!}(y/\mu)^{i-N}(1-y/\mu)^2 + \frac{N^N}{N!}(1-y/\mu)$ is strictly decreasing in y for $y < \mu$, $\widetilde{W}_p(Ny)$ is strictly increasing in y. □

Lemma 19. *Recall Remark A.12.1. In the unobservable dedicated system, the average sojourn time and social welfare in equilibrium are respectively given as*

$$\widehat{W}_d = \begin{cases} \frac{R-f}{c} & \text{if } \rho < 1 \text{ and } \frac{R-f}{c} < \frac{1}{\mu-\lambda}, \text{ or } \rho \geq 1 \\ \frac{1}{\mu-\lambda} & \text{if } \rho < 1 \text{ and } \frac{1}{\mu-\lambda} \leq \frac{R-f}{c}, \end{cases} \tag{A.220}$$

$$\widehat{SW}_d = \begin{cases} \left(\mu - \frac{c}{R-f}\right) fN & \text{if } \rho < 1 \text{ and } \frac{R-f}{c} < \frac{1}{\mu-\lambda}, \text{ or } \rho \geq 1 \\ \left(R\lambda - \frac{c\lambda}{\mu-\lambda}\right) N & \text{if } \rho < 1 \text{ and } \frac{1}{\mu-\lambda} \leq \frac{R-f}{c}. \end{cases} \tag{A.221}$$

Proof of Lemma 19: Each unobservable dedicated queue is an $M/M/1$ queue with service rate μ. Suppose that the effective arrival rate in a queue is $x < \mu$. Then, by Section 7.3.1 of (Kulkarni, 2010), the average sojourn time is $\widetilde{W}_d(x) = \frac{1}{\mu-x}$ and the average number of customers in one of the N separate sub-systems is $\widetilde{L}_d(x) = \frac{x}{\mu-x}$. Based on this,

$$\widetilde{W}_d(\Lambda_d) = \widetilde{W}_d(\lambda) = \begin{cases} \frac{1}{\mu-\lambda} & \text{if } \lambda < \mu \\ \\ \infty & \text{if } \lambda \geq \mu. \end{cases}$$

Using (A.211) through (A.213), the equilibrium effective arrival rate in each queue is

$$\widehat{\lambda}_{e,d} = \begin{cases} \lambda q_d^* = \mu - c/(R-f) & \text{if } \rho < 1 \text{ and } \frac{R-f}{c} < \frac{1}{\mu-\lambda}, \text{ or } \rho \geq 1 \\ \\ \lambda & \text{if } \rho < 1 \text{ and } \frac{1}{\mu-\lambda} \leq \frac{R-f}{c}, \end{cases} \tag{A.222}$$

where $q_d^* = (\mu - c/(R-f))\,\lambda^{-1}$ is the unique solution of $1/(\mu - \lambda q_d^*) = (R-f)/c$, and the equilibrium average sojourn time is

$$\widehat{W}_d = \begin{cases} \frac{R-f}{c} & \text{if } \rho < 1 \text{ and } \frac{R-f}{c} < \frac{1}{\mu-\lambda}, \text{ or } \rho \geq 1 \\ \\ \frac{1}{\mu-\lambda} & \text{if } \rho < 1 \text{ and } \frac{1}{\mu-\lambda} \leq \frac{R-f}{c}. \end{cases}$$

(Note that we did not include the condition $1/\mu < (R-f)/c$ in the first line of (A.222) because (A.210) already implies that.) From Little's law, we have $\widehat{L}_d = \widehat{\lambda}_{e,d}\widehat{W}_d$. Therefore,

$$\widehat{L}_d = \begin{cases} \frac{(R-f)\mu}{c} - 1 & \text{if } \rho < 1 \text{ and } \frac{R-f}{c} < \frac{1}{\mu-\lambda}, \text{ or } \rho \geq 1 \\ \\ \frac{\lambda}{\mu-\lambda} & \text{if } \rho < 1 \text{ and } \frac{1}{\mu-\lambda} \leq \frac{R-f}{c}. \end{cases}$$

Plugging the expressions above in $\widehat{SW}_d$ formula (2.19), we complete the proof of Lemma 19. □

A.12.2 Proof of Proposition 6

Recall Remark A.12.1. Recall also Lemmas 18 and 19, and their proofs. Note that the unobservable pooled system is an $M/M/N$ system with the equilibrium effective arrival rate (A.219) and each unobservable dedicated sub-system (that consists of one dedicated line and its server) is an $M/M/1$ system with the equilibrium effective arrival rate (A.222). Note also that $\widetilde{W}_p(Nx)$ represents the average sojourn time in the $M/M/N$ system with the total effective arrival rate Nx and the service rate μ for each server, and $\widetilde{W}_d(x)$

represents the average sojourn time in the $M/M/1$ system with the effective arrival rate x and service rate μ.

We claim and show in Lemma 20 at the end of this section that

$$\widetilde{W}_p(Nx) \leq \widetilde{W}_d(x) \qquad \text{for } x < \mu, \tag{A.223}$$

and hence

$$\widetilde{W}_p(N\lambda) \leq \widetilde{W}_d(\lambda) \qquad \text{for } \rho < 1. \tag{A.224}$$

Using (A.224), we will prove the claim in Proposition 6 under two main cases. Case 1: Suppose that $\rho < 1$ and $\frac{1}{\mu-\lambda} \leq \frac{R-f}{c}$. Then, by Lemma 19 and its proof, $\widehat{W}_d = \frac{1}{\mu-\lambda}$, $\widehat{q}_d = 1$ and $\widehat{\lambda}_{e,d} = \lambda$ in equilibrium. This and (A.224) imply that $\widetilde{W}_p(N\lambda) \leq \frac{R-f}{c}$. Then, by the proof of Lemma 18, $\widehat{q}_p = 1$ and $\widehat{\lambda}_{e,p} = N\lambda$ in equilibrium. As a result,

$$\widehat{W}_p = \widetilde{W}_p(N\lambda) \leq \widetilde{W}_d(\lambda) = \widehat{W}_d. \tag{A.225}$$

Recall the social welfare from (2.19), and recall that $\widehat{\lambda}_{e,d}N = \widehat{\lambda}_{e,p} = N\lambda$. Then,

$$\widehat{SW}_d = (R - c\widehat{W}_d)\widehat{\lambda}_{e,d}N = (R - c\widehat{W}_d)\lambda N \quad \text{and} \quad \widehat{SW}_p = (R - c\widehat{W}_p)\widehat{\lambda}_{e,p} = (R - c\widehat{W}_p)\lambda N.$$

Because $\widehat{W}_p \leq \widehat{W}_d$ by (A.225), $\widehat{SW}_p \geq \widehat{SW}_d$. This completes the proof of Proposition 6 under Case 1. Case 2: Suppose now that either $\rho < 1$ and $\frac{R-f}{c} < \frac{1}{\mu-\lambda}$, or $\rho \geq 1$. From Lemma 19, it follows that, in equilibrium, the average sojourn time in the unobservable dedicated system is

$$\widehat{W}_d = \frac{R-f}{c}, \tag{A.226}$$

and the equilibrium social welfare in the dedicated system is

$$\widehat{SW}_d = (R - c\widehat{W}_d)\widehat{\lambda}_{e,d}N = f\widehat{\lambda}_{e,d}N. \tag{A.227}$$

Given these performance metrics in the dedicated system, we now prove the claim by considering the following two subcases for $\widetilde{W}_p(N\lambda)$.

Case 2.1: Suppose that $\widetilde{W}_p(N\lambda) \geq \frac{R-f}{c}$. Then, by Lemma 18, the equilibrium average sojourn time in the pooled system is $\widehat{W}_p = \frac{R-f}{c}$, which is equal to $\widehat{W}_d$ by (A.226). Thus, by Lemma 18,

$$\widehat{SW}_p = (R - c\widehat{W}_p)\widehat{\lambda}_{e,p} = f\widehat{\lambda}_{e,p}. \tag{A.228}$$

We now show that

$$\widehat{\lambda}_{e,p} \geq \widehat{\lambda}_{e,d}N. \tag{A.229}$$

Suppose for a contradiction that $\widehat{\lambda}_{e,p} < N\widehat{\lambda}_{e,d}$. Then, (A.223) and the fact that $\widetilde{W}_p(x)$ strictly increases in x for $x < N\mu$ imply that

$$\widehat{W}_p = \widetilde{W}_p(\widehat{\lambda}_{e,p}) < \widetilde{W}_p(N\widehat{\lambda}_{e,d}) \leq \widetilde{W}_d(\widehat{\lambda}_{e,d}) = \widehat{W}_d. \tag{A.230}$$

But, this contradicts with $\widehat{W}_p = \widehat{W}_d$. Thus, we have (A.229). Based on (A.229), from (A.227) and (A.228), it follows that $\widehat{SW}_p \geq \widehat{SW}_d$.

Case 2.2: Suppose that $\widetilde{W}_p(N\lambda) < \frac{R-f}{c}$. Then, $\widehat{q}_p = 1$ and $\widehat{\lambda}_{e,p} = N\lambda$ in equilibrium. Thus, $R-f-c\widehat{W}_p = R - f - c\widetilde{W}_p(N\lambda) > 0$, which implies that the equilibrium long-run average sojourn time in the pooled system satisfies $\widehat{W}_p < \frac{R-f}{c} = \widehat{W}_d$. Thus, the equilibrium social welfare in the pooled system is

$$\widehat{SW}_p = (R - c\widehat{W}_p)\widehat{\lambda}_{e,p} = (R - c\widehat{W}_p)N\lambda > (R - c\widehat{W}_d)N\lambda \geq (R - c\widehat{W}_d)\widehat{\lambda}_{e,d}N = \widehat{SW}_d,$$

which completes the proof of Case 2.2.

Note that combining Case 1 and Case 2 covers the entire parameter space. Thus, the claim in Proposition 6 follows.

The following lemma shows our claim in (A.223).

Lemma 20. $\widetilde{W}_p(Nx) \leq \widetilde{W}_d(x)$ *for* $x/\mu < 1$.

Proof of Lemma 20: Suppose that $x/\mu < 1$. Recall that $\widetilde{W}_p(Nx)$ represents the average sojourn time in the $M/M/N$ system with the total effective arrival rate Nx and the service rate μ for each server, and $\widetilde{W}_d(x)$ represents the average sojourn time in the $M/M/1$ system with the effective arrival rate x and service rate

μ. Denote by X_d total number of customers in N of the $M/M/1$ lines in the steady-state, and let X_p be the corresponding figure in the aforementioned $M/M/N$ system. Based on this, to show our claim, we will use standard likelihood comparison technique (see, for instance, (Smith and Whitt, 1981)). Let $\theta_p(m+1)$ be the transition rate from state $m+1$ to m in the pooled system (in the steady-state), $\theta_d(m+1|S_t)$ be the transition rate from state $m+1$ to m in the dedicated system and S_t is the state of the dedicated system (i.e., number of customers in each of the N lines) at time t, for any $m = 0, 1, \ldots$. Because $\theta_d(m+1|S_t) \leq \theta_p(m+1)$ for each m regardless of S_t, in the steady-state, we have

$$\mathbb{P}(X_d = m)Nx \leq \mathbb{P}(X_d = m+1)\theta_p(m+1) \quad \text{and} \quad \mathbb{P}(X_p = m)Nx = \mathbb{P}(X_p = m+1)\theta_p(m+1).$$

Thus, we have

$$\frac{\mathbb{P}(X_d = m+1)}{\mathbb{P}(X_d = m)} \geq \frac{Nx}{\theta_p(m+1)} = \frac{\mathbb{P}(X_p = m+1)}{\mathbb{P}(X_p = m)}. \tag{A.231}$$

Using this, we now show that $\mathbb{E}(X_d) \geq \mathbb{E}(X_p)$. Note that (A.231) implies that $\frac{\mathbb{P}(X_d=j)}{\mathbb{P}(X_d=i)} \geq \frac{\mathbb{P}(X_p=j)}{\mathbb{P}(X_p=i)}$ for all $i \leq j$, $i, j \in \mathbb{N}$, which is equivalent to

$$\mathbb{P}(X_d = j)\mathbb{P}(X_p = i) \geq \mathbb{P}(X_d = i)\mathbb{P}(X_p = j). \tag{A.232}$$

The summation on both sides of (A.232) over i from 0 to j gives

$$\mathbb{P}(X_d = j)\mathbb{P}(X_p \leq j) \geq \mathbb{P}(X_d \leq j)\mathbb{P}(X_p = j). \tag{A.233}$$

Similarly, the summation on both sides of (A.232) over j from $i+1$ to ∞ results in

$$\mathbb{P}(X_d \geq i+1)\mathbb{P}(X_p = i) \geq \mathbb{P}(X_d = i)\mathbb{P}(X_p \geq i+1). \tag{A.234}$$

Combining (A.233) and (A.234) and letting $i = j = a$, we have

$$\frac{\mathbb{P}(X_d \geq a+1)}{\mathbb{P}(X_p \geq a+1)} \geq \frac{\mathbb{P}(X_d = a)}{\mathbb{P}(X_p = a)} \geq \frac{\mathbb{P}(X_d \leq a)}{\mathbb{P}(X_p \leq a)}. \tag{A.235}$$

Thus, $\mathbb{P}(X_d \leq a) \leq \mathbb{P}(X_p \leq a)$ for any non-negative integer a, and hence

$$\mathbb{E}(X_d) = \sum_{i=0}^{\infty}(1 - \mathbb{P}(X_d \leq i)) \geq \sum_{i=0}^{\infty}(1 - \mathbb{P}(X_p \leq i)) = \mathbb{E}(X_p). \tag{A.236}$$

Observe that the long-run average number of customers in one of the N separate dedicated sub-systems (i.e., $\widetilde{L}_d$) and the long-run average number of customers in the pooled system (i.e., $\widetilde{L}_p$) satisfy $\widetilde{L}_d = \mathbb{E}(X_d)/N$ and $\widetilde{L}_p \doteq \mathbb{E}(X_p)$. Then, by Little's Law and (A.236),

$$\widetilde{W}_d(x) = \frac{\widetilde{L}_d(x)}{x} = \frac{\mathbb{E}(X_d)/N}{x} = \frac{\mathbb{E}(X_d)}{Nx} \geq \frac{\mathbb{E}(X_p)}{Nx} = \frac{\widetilde{L}_p(Nx)}{Nx} = \widetilde{W}_p(Nx).$$

This completes the proof of the claim. □ □

A.13 Proof of Proposition 7

Suppose that the reward has a general distribution with the p.d.f. $g(\cdot)$ and the c.d.f $G(\cdot)$ defined on the support $[L, H]$, and consider unobservable systems. We prove the claim under each of the three possible cases about $c\widetilde{W}_d(\cdot)$ below:

Case 1: Suppose that $c\widetilde{W}_d(\lambda) \leq L$. Then, $\lambda < \mu$ must be true and in equilibrium all customers join the dedicated system. By Lemma 20, $c\widetilde{W}_p(N\lambda) \leq c\widetilde{W}_d(\lambda) \leq L$, thus all customers in the pooled system also join. As a result, in equilibrium, $\widehat{W}_p = \widetilde{W}_p(N\lambda) \leq \widetilde{W}_d(\lambda) = \widehat{W}_d$ and $\widehat{SW}_p = N\lambda \int_L^H g(x)(x - c\widehat{W}_p)dx \geq N\lambda \int_L^H g(x)(x - c\widehat{W}_d)dx = \widehat{SW}_d$.

Case 2: Suppose that $c\widetilde{W}_d(0) = \frac{c}{\mu} > H$. Then, no customer in the dedicated system joins. Moreover, no customer joins in the pooled system as well since $c\widetilde{W}_p(0) = \frac{c}{\mu} > H$.

Case 3: Suppose that $c\widetilde{W}_d(\lambda) > L$ and $c\widetilde{W}_d(0) \leq H$. Then, there exits a threshold reward a_d^e such that only customers with reward larger than a_d^e join in equilibrium and $a_d^e = c\widetilde{W}_d(\lambda\bar{G}(a_d^e))$ where $\bar{G}(x) \doteq 1 - G(x)$. Note that a_d^e is the intersection point of functions $y_1(x) \doteq x$ and $y_2(x) \doteq c\widetilde{W}_d(\lambda\bar{G}(x))$ for $x \in [L, H]$. (The intersection point exists and is unique as $y_1(\cdot)$ is strictly increasing and $y_2(\cdot)$ is strictly decreasing.) We consider two possible subcases: Case 3.1: Suppose that $c\widetilde{W}_p(N\lambda) \leq L$. Then, in the pooled system, all customers join. Because $c\widetilde{W}_p(N\lambda) \leq L < a_d^e = c\widetilde{W}_d(\lambda\bar{G}(a_d^e))$, $\widetilde{W}_p(N\lambda) < \widetilde{W}_d(\lambda\bar{G}(a_d^e))$, i.e., $\widehat{W}_p < \widehat{W}_d$, and $\widehat{SW}_p = N\lambda \int_L^H g(x)(x - c\widehat{W}_p)dx > N\lambda \int_{a_d^e}^H g(x)(x - c\widehat{W}_p)dx > N\lambda \int_{a_d^e}^H g(x)(x - c\widehat{W}_d)dx = \widehat{SW}_d$. Case 3.2: Suppose that $c\widetilde{W}_p(N\lambda) > L$ and $c\widetilde{W}_p(0) \leq H$. Then, there exits a threshold

reward a_p^e such that only customers with reward larger than a_p^e join in equilibrium and $a_p^e = c\widetilde{W}_p(N\lambda\bar{G}(a_p^e))$. Note that a_p^e is the intersection point of functions $y_1(x)$ and $y_3(x) \doteq c\widetilde{W}_p(N\lambda\bar{G}(x))$ for $x \in [L, H]$. Since $y_3(x) = c\widetilde{W}_p(N\lambda\bar{G}(x)) \leq c\widetilde{W}_d(\lambda\bar{G}(x)) = y_2(x)$ for any $x \in [L, H]$, $a_p^e \leq a_d^e$ and $\widetilde{W}_p(N\lambda\bar{G}(a_p^e)) \leq \widetilde{W}_d(\lambda\bar{G}(a_d^e))$, i.e., $\widehat{W}_p \leq \widehat{W}_d$. As a result, $\widehat{SW}_p = N\lambda\int_{a_p^e}^{H} g(x)(x - c\widehat{W}_p)dx \geq N\lambda\int_{a_d^e}^{H} g(x)(x - c\widehat{W}_p)dx \geq N\lambda\int_{a_d^e}^{H} g(x)(x - c\widehat{W}_d)dx = \widehat{SW}_d$. □

A.13.1 Proof of Proposition 8

Proof of Proposition 8 - Part (a): Recall that the maximum social welfare under the welfare maximization formulation is as in (2.20). Note that (2.20), i.e., choosing the fee to maximize equilibrium social welfare, is equivalent to choosing the effective arrival rate to maximize equilibrium social welfare as below:

$$\widehat{SW}_j^* = \begin{cases} \max_{0\leq x\leq\lambda}(R - c\widetilde{W}_j(x))xN & \text{if } j = d, \\ \max_{0\leq x\leq\lambda}(R - c\widetilde{W}_j(Nx))xN & \text{if } j = p. \end{cases} \tag{A.237}$$

We only need to focus on the case that $x < \mu$ in both systems since the the system will be unstable otherwise. We already know that $\widetilde{W}_p(Nx) \leq \widetilde{W}_d(x)$ for $x/\mu < 1$ according to Lemma 20. Thus, $\widehat{SW}_p^* \geq \widehat{SW}_d^*$.□

Proof of Proposition 8 - Part (b): Note that (2.21) is equivalent to choosing the effective arrival rate to maximize the equilibrium revenue

$$\widehat{RV}_j^{**} = \begin{cases} \max_{0\leq x\leq\lambda}(R - c\widetilde{W}_j(x))xN & \text{if } j = d, \\ \max_{0\leq x\leq\lambda}(R - c\widetilde{W}_j(Nx))xN & \text{if } j = p. \end{cases} \tag{A.238}$$

This and (A.237) imply that the maximum revenue is the same as the maximum social welfare in equilibrium. Then, the claim in part (b) immediately follows from part (a). □

A.14 Proof of Proposition 9

A.14.1 Proof of Part (a)

We already proved in Proposition 6 that the unobservable pooled system outperforms the unobservable dedicated system in social welfare. Based on this, we only need to compare the unobservable pooled system

with the observable pooled system to prove Proposition 9-(a). Below, we will show that the observable pooled system results in larger equilibrium social welfare than the unobservable pooled system. There can be two cases related to $\widetilde{W}_p(N\lambda)$:

Case A: Suppose that $\widetilde{W}_p(N\lambda) > \frac{R}{c}$. Then, by Lemma 18, in the unobservable pooled system, average sojourn time and social welfare in equilibrium are the following, respectively:

$$\widehat{W}_p = \frac{R}{c} \quad \text{and} \quad \widehat{SW}_p = (R - c\widehat{W}_p)\widehat{\lambda}_{e,p} = 0. \tag{A.239}$$

Recall the social welfare SW_p in the observable pooled system from Lemma 1. We claim and show below that $SW_p > 0$. Combining this with (A.239), we have $\widehat{SW}_p < SW_p$.

It only remains to prove our claim that $SW_p > 0$. Recall from (A.210) that $\frac{R\mu}{c} > 1$. Then, each joining customer receives a non-negative expected net benefit by joining. Furthermore, a customer that finds $n \leq N-1$ customers in the system upon arrival receives strictly positive expected net benefit $R - \bar{W}_p(n+1)c = R - \frac{1}{\mu}c > 0$ for $\frac{R\mu}{c} > 1$. As a result, $SW_p > 0$ in this case.

Case B: Suppose that $\widetilde{W}_p(N\lambda) \leq \frac{R}{c}$. Then, by Lemma 18 and its proof, $\widehat{q}_p = 1$ and $\widehat{\lambda}_{e,p} = N\lambda$ for the unobservable pooled system. We now show that the observable pooled system results in strictly larger social welfare than the unobservable pooled system, i.e., $SW_p > \widehat{SW}_p$. Recall the SW_p from Lemma 1, and observe that using the stationary probability distribution $\{\pi_0, \pi_1, \ldots, \pi_K\}$ for the number of customers

in the observable pooled system (in the proof of Lemma 1), SW_p can also be expressed as follows:

$$\begin{aligned}
SW_p =& N\lambda \left(\sum_{i=0}^{N-1} \left(R - \frac{c}{\mu} \right) \frac{N^i}{i!}\rho^i + \sum_{i=N}^{K-1} \left(R - \frac{(i+1)c}{N\mu} \right) \frac{N^N}{N!}\rho^i \right) \Big/ \left(\sum_{i=0}^{N-1} \frac{N^i}{i!}\rho^i + \sum_{i=N}^{K} \frac{N^N}{N!}\rho^i \right) \\
=& N\lambda \left(\sum_{i=0}^{N-1} (R - \bar{W}_p(i+1)c) \frac{N^i}{i!}\rho^i + \sum_{i=N}^{K-1} (R - \bar{W}_p(i+1)c) \frac{N^N}{N!}\rho^i \right) \Big/ \left(\sum_{i=0}^{N-1} \frac{N^i}{i!}\rho^i + \sum_{i=N}^{K} \frac{N^N}{N!}\rho^i \right) \\
>& N\lambda \left(\sum_{i=0}^{N-1} (R - \bar{W}_p(i+1)c) \frac{N^i}{i!}\rho^i + \sum_{i=N}^{\infty} (R - \bar{W}_p(i+1)c) \frac{N^N}{N!}\rho^i \right) \Big/ \left(\sum_{i=0}^{N-1} \frac{N^i}{i!}\rho^i + \sum_{i=N}^{\infty} \frac{N^N}{N!}\rho^i \right) \qquad \text{(A.240)} \\
=& N\lambda R - cN\lambda \left(\sum_{i=0}^{N-1} \bar{W}_p(i+1) \frac{N^i}{i!}\rho^i + \sum_{i=N}^{\infty} \bar{W}_p(i+1) \frac{N^N}{N!}\rho^i \right) \Big/ \left(\sum_{i=0}^{N-1} \frac{N^i}{i!}\rho^i + \sum_{i=N}^{\infty} \frac{N^N}{N!}\rho^i \right) \\
=& N\lambda R - cN\lambda \left(\sum_{i=0}^{N-1} \frac{1}{\mu} \frac{N^i}{i!}\rho^i + \sum_{i=N}^{\infty} \frac{i+1}{N\mu} \frac{N^N}{N!}\rho^i \right) \Big/ \left(\sum_{i=0}^{N-1} \frac{N^i}{i!}\rho^i + \sum_{i=N}^{\infty} \frac{N^N}{N!}\rho^i \right) \\
=& N\lambda R - c \left(\sum_{i=0}^{N-1} \frac{N^{i+1}}{i!}\rho^{i+1} + \sum_{i=N}^{\infty} \frac{N^N}{N!}(i+1)\rho^{i+1} \right) \Big/ \left(\sum_{i=0}^{N-1} \frac{N^i}{i!}\rho^i + \sum_{i=N}^{\infty} \frac{N^N}{N!}\rho^i \right) \\
=& N\lambda R - c \left(\sum_{i=0}^{N-1} \frac{N^i}{i!} i\rho^i + \sum_{i=N}^{\infty} \frac{N^N}{N!} i\rho^i \right) \Big/ \left(\sum_{i=0}^{N-1} \frac{N^i}{i!}\rho^i + \sum_{i=N}^{\infty} \frac{N^N}{N!}\rho^i \right) \qquad \text{(A.241)} \\
=& \widehat{SW}_p
\end{aligned}$$

Here, the inequality (A.240) holds because $R - \bar{W}_p(i+1)c = R - \frac{i+1}{N\mu}c < 0$ for any $i \geq K \doteq \lfloor \frac{RN\mu}{c} \rfloor$.

Combining Cases A and B, it follows that the observable pooled systems results in larger social welfare than the unobservable pooled system. This and Proposition 6 complete our proof for Proposition 9-(a). □

A.14.2 Proof of Part (b)

According to Propositions 4 and 8, when the service fee is set to maximize the social welfare, the maximum social welfare in the pooled system is larger than that in the dedicated system for both observable and unobservable cases. Thus, to prove the claim, we only need to compare the observable pooled system with the unobservable pooled system. To do so, consider an alternative setting in which admissions to an $M/M/N$ system can be controlled rather than customers making their own joining/balking decisions. Among all admission control policies (including the randomized ones), the optimal admission rule that maximizes the social welfare is a deterministic control limit rule that induces a queue capacity. (This is because this alternative formulation corresponds to a finite state Markov decision process.) Because that

optimal queue capacity can be achieved by imposing a service fee in the observable pooled system where customers make their own joining decisions, the observable pooled system achieves larger maximum social welfare than the unobservable pooled system when in each system, the fee is set to maximize the social welfare. From this, the claim immediately follows. □

A.15 Proof of Lemma 4

We will first state and prove three lemmas, statements of which will be used in the remainder of the proof.

Lemma 21. *When the service fee is set to maximize the revenue, there exists a constant $\bar{\lambda}_1$ such that if the arrival rate parameter λ satisfies $\lambda < \bar{\lambda}_1$, then the maximum revenue in the unobservable pooled system is achieved with the effective arrival rate $N\lambda$.*

Proof of Lemma 21: By (A.210), we only need to consider the case that $R - \frac{c}{\mu} > 0$, which implies $\widetilde{W}_p(0) = \frac{1}{\mu} < \frac{R}{c}$. From the proof of Lemma 18, we already know that $\widetilde{W}_p(Ny)$ is a continuous function for $y < \mu$ and it is strictly increasing in y. Thus, there exists $\bar{\lambda}_0$ such that $\widetilde{W}_p(N\lambda) \leq \frac{R}{c}$ for $\lambda < \bar{\lambda}_0$. Then, as a function of effective arrival rate Nx for $0 \leq x < \mu$, the revenue in the unobservable pooled system is

$$\begin{aligned}
\widehat{RV}_p(Nx) &= (R - c\widetilde{W}_p(Nx))Nx \\
&= NRx - \left(\frac{\sum_{i=0}^{N-1} \frac{N^i}{i!} i(x/\mu)^i + \frac{N^N}{N!} \sum_{i=N}^{\infty} i(x/\mu)^i}{\left(\sum_{i=0}^{N-1} \frac{N^i}{i!} (x/\mu)^i + \frac{N^N}{N!} \sum_{i=N}^{\infty} (x/\mu)^i \right)} \right) c \\
&= NRx - \left(\frac{\sum_{i=1}^{N} \frac{N^i}{i!} i(x/\mu)^i + \frac{N^N}{N!} \sum_{i=N+1}^{\infty} i(x/\mu)^i}{\left(\sum_{i=0}^{N-1} \frac{N^i}{i!} (x/\mu)^i + \frac{N^N}{N!} \sum_{i=N}^{\infty} (x/\mu)^i \right)} \right) c \\
&= NRx - \left(\frac{\sum_{i=0}^{N-1} \frac{N^i}{i!} (x/\mu)^i Nx/\mu + \frac{N^N}{N!} \sum_{i=N}^{\infty} (i+1)(x/\mu)^{i+1}}{\left(\sum_{i=0}^{N-1} \frac{N^i}{i!} (x/\mu)^i + \frac{N^N}{N!} \sum_{i=N}^{\infty} (x/\mu)^i \right)} \right) c \\
&= NRx - \frac{Nx}{\mu} c - \frac{\frac{N^N}{N!} \sum_{i=N}^{\infty} (i+1-N)(x/\mu)^{i+1}}{\left(\sum_{i=0}^{N-1} \frac{N^i}{i!} (x/\mu)^i + \frac{N^N}{N!} \sum_{i=N}^{\infty} (x/\mu)^i \right)} c.
\end{aligned}$$

Thus,

$$
\begin{aligned}
&\frac{d\widehat{RV}_p(Nx)}{dx}\\
=&NR-\frac{Nc}{\mu}-\frac{\left(\frac{N^N}{N!}\sum_{i=N}^{\infty}(i+1-N)(i+1)(x/\mu)^i\frac{1}{\mu}\right)\left(\sum_{i=0}^{N-1}\frac{N^i}{i!}(x/\mu)^i+\frac{N^N}{N!}\sum_{i=N}^{\infty}(x/\mu)^i\right)}{\left(\sum_{i=0}^{N-1}\frac{N^i}{i!}(x/\mu)^i+\frac{N^N}{N!}\sum_{i=N}^{\infty}(x/\mu)^i\right)^2}c\\
+&\frac{\left(\frac{N^N}{N!}\sum_{i=N}^{\infty}(i+1-N)(x/\mu)^{i+1}\right)\left(\sum_{i=1}^{N-1}\frac{N^i}{i!}i(x/\mu)^{i-1}\frac{1}{\mu}+\frac{N^N}{N!}\sum_{i=N}^{\infty}i(x/\mu)^{i-1}\frac{1}{\mu}\right)}{\left(\sum_{i=0}^{N-1}\frac{N^i}{i!}(x/\mu)^i+\frac{N^N}{N!}\sum_{i=N}^{\infty}(x/\mu)^i\right)^2}c\\
>&NR-\frac{Nc}{\mu}-\frac{\frac{N^N}{N!}\sum_{i=N}^{\infty}(i+1-N)(i+1)(x/\mu)^i\frac{1}{\mu}}{\left(\sum_{i=0}^{N-1}\frac{N^i}{i!}(x/\mu)^i+\frac{N^N}{N!}\sum_{i=N}^{\infty}(x/\mu)^i\right)}c\\
>&NR-\frac{Nc}{\mu}-\frac{\frac{N^N}{N!}\sum_{i=N}^{\infty}(i+1)^2(x/\mu)^i\frac{1}{\mu}}{\left(\sum_{i=0}^{N-1}\frac{N^i}{i!}(x/\mu)^i+\frac{N^N}{N!}\sum_{i=N}^{\infty}(x/\mu)^i\right)}c\\
\geq&NR-\frac{Nc}{\mu}-\frac{\frac{N^N}{N!}\sum_{i=1}^{\infty}(i+1)^2(x/\mu)^i\frac{1}{\mu}}{\left(\sum_{i=0}^{N-1}\frac{N^i}{i!}(x/\mu)^i+\frac{N^N}{N!}\sum_{i=N}^{\infty}(x/\mu)^i\right)}c\\
>&NR-\frac{Nc}{\mu}-\frac{N^N}{N!}\sum_{i=2}^{\infty}i^2(x/\mu)^{i-1}\frac{c}{\mu}\\
=&NR-\frac{Nc}{\mu}-\frac{N^N}{N!}\frac{((x/\mu)^2-3(x/\mu)+4)(x/\mu)}{(1-x/\mu)^3}\frac{c}{\mu}.
\end{aligned}
$$

Recall that $NR-\frac{Nc}{\mu}>0$. Because $\frac{((x/\mu)^2-3(x/\mu)+4)(x/\mu)}{(1-x/\mu)^3}\frac{c}{\mu}$ is continuous and increasing in x, and $\lim_{x\to 0}\frac{N^N}{N!}\frac{((x/\mu)^2-3(x/\mu)+4)(x/\mu)}{(1-x/\mu)^3}\frac{c}{\mu}=0$, there exists $\bar{\lambda}_1$ such that $\bar{\lambda}_1<\bar{\lambda}_0$ and for $x<\bar{\lambda}_1$,

$$
\frac{N^N}{N!}\frac{((x/\mu)^2-3(x/\mu)+4)(x/\mu)}{(1-x/\mu)^3}\frac{c}{\mu}<NR-\frac{Nc}{\mu}. \tag{A.242}
$$

As a result, $\frac{d\widehat{RV}_p(Nx)}{dx}>0$ for $x<\bar{\lambda}_1$. This implies that when the market size parameter $\lambda<\bar{\lambda}_1$, $\widehat{RV}_p(Nx)$ is increasing in x and the revenue is maximized when $Nx=N\lambda$. □

Lemma 22. *When the service fee is set to maximize revenue, there exists a constant $\bar{\lambda}_2$ such that if the arrival rate parameter $\lambda<\bar{\lambda}_2$, then the optimal service fee in the observable pooled system is $f_p^{**}=R-\frac{c}{\mu}$ and the corresponding system capacity is $K_p^{**}=N$.*

Proof of Lemma 22: Consider any fixed balking threshold (i.e., capacity) K such that $N \leq K \leq \lfloor \frac{RN\mu}{c} \rfloor$. Then, the revenue in the observable pooled system as a function of capacity K is

$$RV_p(K) = N\lambda(1 - b_p(K)) \left(R - \frac{cK}{N\mu} \right),$$

where $b_p(K)$ is the balking probability in the observable pooled system when the capacity is K. Thus,

$$RV_p'(K) = N\lambda \left(-b_p'(K) \left(R - \frac{cK}{N\mu} \right) - (1 - b_p(K)) \frac{c}{N\mu} \right).$$

Because by (A.13), the balking probability is

$$b_p(K) = \frac{\frac{N^N}{N!}\rho^K}{\sum_{i=0}^{N-1} \frac{N^i}{i!}\rho^i + \sum_{i=N}^{K} \frac{N^N}{N!}\rho^i},$$

$$b_p'(K) = \frac{\frac{N^N}{N!}\rho^K \ln(\rho)}{\sum_{i=0}^{N-1} \frac{N^i}{i!}\rho^i + \sum_{i=N}^{K} \frac{N^N}{N!}\rho^i} - \frac{\frac{N^N}{N!}\rho^K \frac{N^N}{N!} \frac{\rho^{K+1}\ln(\rho)}{\rho - 1}}{(\sum_{i=0}^{N-1} \frac{N^i}{i!}\rho^i + \sum_{i=N}^{K} \frac{N^N}{N!}\rho^i)^2} < 0.$$

Thus,

$$b_p(K) \leq b_p(N) = \frac{\frac{N^N}{N!}\rho^N}{\sum_{i=0}^{N} \frac{N^i}{i!}\rho^i}. \tag{A.243}$$

Note that when $\rho < 1$ and $K \geq N$,

$$-b_p'(K) \leq -\frac{\frac{N^N}{N!}\rho^N \ln(\rho)}{\sum_{i=0}^{N} \frac{N^i}{i!}\rho^i} + \frac{\frac{N^N}{N!}\rho^N \frac{N^N}{N!} \frac{\rho^{N+1}\ln(\rho)}{\rho - 1}}{(\sum_{i=0}^{N} \frac{N^i}{i!}\rho^i)^2},$$

$b_p(N)$ and $-\frac{\frac{N^N}{N!}\rho^N \ln(\rho)}{\sum_{i=0}^{N} \frac{N^i}{i!}\rho^i} + \frac{\frac{N^N}{N!}\rho^N \frac{N^N}{N!} \frac{\rho^{N+1}\ln(\rho)}{\rho - 1}}{(\sum_{i=0}^{N} \frac{N^i}{i!}\rho^i)^2}$ are both continuous in ρ, and satisfy the following relations:

$$\begin{aligned} &\lim_{\rho \to 0} b_p(N) = \lim_{\rho \to 0} \frac{\frac{N^N}{N!}\rho^N}{\sum_{i=0}^{N} \frac{N^i}{i!}\rho^i} = 0 \\ &\lim_{\rho \to 0} -\frac{\frac{N^N}{N!}\rho^N \ln(\rho)}{\sum_{i=0}^{N} \frac{N^i}{i!}\rho^i} + \frac{\frac{N^N}{N!}\rho^N \frac{N^N}{N!} \frac{\rho^{N+1}\ln(\rho)}{\rho - 1}}{(\sum_{i=0}^{N} \frac{N^i}{i!}\rho^i)^2} = 0. \end{aligned} \tag{A.244}$$

Thus, there exists a constant $\bar{\lambda}_2 < \mu$ such that $\lambda < \bar{\lambda}_2$ and $K \geq N$ imply $-b_p'(K) < \frac{1}{2}\frac{c}{(R\mu - c)N}$ and $b_p(K) \leq b_p(N) < \frac{1}{2}$. As a result, when $\lambda < \bar{\lambda}_2$ and $K \geq N$,

$$\begin{aligned} RV_p'(K) &= N\lambda \left(-b_p'(K) \left(R - \frac{cK}{N\mu} \right) - (1 - b_p(K)) \frac{c}{N\mu} \right) \\ &< N\lambda \left(\frac{1}{2} \frac{c}{(R\mu - c)N} \left(R - \frac{cN}{N\mu} \right) - \frac{1}{2} \frac{c}{N\mu} \right) = 0, \end{aligned}$$

which implies that $RV_p(K)$ achieves the maximum with the fee $f_p^{**} = R - \frac{c}{\mu}$ and the corresponding capacity $K_p^{**} = N$. $\square$

Lemma 23. *When the fee is set to maximize the revenue, there exists a constant $\bar{\lambda}$ such that if $\frac{R\mu}{c} > \frac{N+1}{N}$ and $\lambda < \bar{\lambda}$, the social welfare in the unobservable pooled system is strictly larger than that in the observable pooled system, i.e., $\widehat{SW}_p^{**} > SW_p^{**}$.*

Proof of Lemma 23: According to Lemmas 21 and 22, when the service fee is set to maximize the revenue and $\lambda < \min\{\bar{\lambda}_1, \bar{\lambda}_2\} < \mu$, the social welfare of the unobservable pooled system and the social welfare of the observable pooled system are respectively as follows.

$$\widehat{SW}_p^{**} = N\lambda R - \left(\frac{\sum_{i=0}^{N-1} \frac{N^i}{i!} i\rho^i + \frac{N^N}{N!} \sum_{i=N}^{\infty} i\rho^i}{\sum_{i=0}^{N-1} \frac{N^i}{i!} \rho^i + \frac{N^N}{N!} \sum_{i=N}^{\infty} \rho^i} \right) c. \tag{A.245}$$

$$SW_p^{**} = N\lambda R \left(1 - \frac{\frac{N^N}{N!}\rho^N}{\sum_{i=0}^{N} \frac{N^i}{i!}\rho^i} \right) - \left(\frac{\sum_{i=0}^{N} \frac{N^i}{i!} i\rho^i}{\sum_{i=0}^{N} \frac{N^i}{i!}\rho^i} \right) c. \tag{A.246}$$

Then,

$$\begin{aligned}
\widehat{SW}_p^{**} - SW_p^{**} &= N\lambda R\frac{\frac{N^N}{N!}\rho^N}{\sum_{i=0}^{N}\frac{N^i}{i!}\rho^i} - \left(\frac{\sum_{i=0}^{N-1}\frac{N^i}{i!}i\rho^i + \frac{N^N}{N!}\sum_{i=N}^{\infty} i\rho^i}{\sum_{i=0}^{N-1}\frac{N^i}{i!}\rho^i + \frac{N^N}{N!}\sum_{i=N}^{\infty}\rho^i} - \frac{\sum_{i=0}^{N}\frac{N^i}{i!}i\rho^i}{\sum_{i=0}^{N}\frac{N^i}{i!}\rho^i}\right)c \\
&> N\lambda R\frac{\frac{N^N}{N!}\rho^N}{\sum_{i=0}^{N}\frac{N^i}{i!}\rho^i} - \left(\frac{\sum_{i=0}^{N-1}\frac{N^i}{i!}i\rho^i + \frac{N^N}{N!}\sum_{i=N}^{\infty} i\rho^i}{\sum_{i=0}^{N}\frac{N^i}{i!}\rho^i} - \frac{\sum_{i=0}^{N}\frac{N^i}{i!}i\rho^i}{\sum_{i=0}^{N}\frac{N^i}{i!}\rho^i}\right)c \\
&= N\lambda R\frac{\frac{N^N}{N!}\rho^N}{\sum_{i=0}^{N}\frac{N^i}{i!}\rho^i} - \left(\frac{\frac{N^N}{N!}\sum_{i=N+1}^{\infty} i\rho^i}{\sum_{i=0}^{N}\frac{N^i}{i!}\rho^i}\right)c \\
&= \frac{\frac{N^N}{N!}\rho^N}{\sum_{i=0}^{N}\frac{N^i}{i!}\rho^i}c\left(\frac{N\lambda R}{c} - \frac{\sum_{i=N+1}^{\infty} i\rho^i}{\rho^N}\right) \\
&= \frac{\frac{N^N}{N!}\rho^N}{\sum_{i=0}^{N}\frac{N^i}{i!}\rho^i}c\left(\frac{N\lambda R}{c} - \frac{\frac{(N+1)\rho^{N+1} - N\rho^{N+2}}{(1-\rho)^2}}{\rho^N}\right) \\
&= \frac{\frac{N^N}{N!}\rho^N}{\sum_{i=0}^{N}\frac{N^i}{i!}\rho^i}c\left(\frac{N\lambda R}{c} - \frac{(N+1)\rho - N\rho^2}{(1-\rho)^2}\right) \\
&> \frac{\frac{N^N}{N!}\rho^N}{\sum_{i=0}^{N}\frac{N^i}{i!}\rho^i}c\rho\left(\frac{NR\mu}{c} - \frac{(N+1)}{(1-\rho)^2}\right). \qquad \text{(A.247)}
\end{aligned}$$

If $\frac{R\mu}{c} > \frac{N+1}{N}$,

$$\lim_{\lambda\to 0}\frac{NR\mu}{c} - \frac{(N+1)}{(1-\rho)^2} = \frac{NR\mu}{c} - (N+1) > 0.$$

Because $\frac{NR\mu}{c} - \frac{(N+1)}{(1-\rho)^2}$ is continuous and decreasing in λ, when $\frac{R\mu}{c} > \frac{N+1}{N}$, there exists a constant $\bar{\lambda}_3$ such that $\frac{NR\mu}{c} - \frac{(N+1)}{(1-\rho)^2} > 0$ for $\lambda < \bar{\lambda}_3$. Define $\bar{\lambda} \doteq \min\{\bar{\lambda}_1, \bar{\lambda}_2, \bar{\lambda}_3\}$. Then, it follows from (A.247) that $\widehat{SW}_p^{**} > SW_p^{**}$ when $\lambda < \bar{\lambda}$ and $\frac{R\mu}{c} > \frac{N+1}{N}$. □

Define $\widehat{\eta}$ as

$$\widehat{\eta} \doteq \left(\frac{(1+\rho)^2}{\rho} + 1\right)\frac{1}{\mu}.$$

Recall also the notation $RV_d(\cdot)$ from the proof of Lemma 3. Suppose that $R/c < \widehat{\eta}$. Note that $R/c < \widehat{\eta}$ if and only if

$$\begin{aligned}
R\rho < (1+\rho^2+3\rho)\frac{c}{\mu} &\Longleftrightarrow N\lambda\left(\frac{1}{1+\rho}\right)\left(R - \frac{c}{\mu}\right) > N\lambda\left(\frac{1+\rho}{1+\rho+\rho^2}\right)\left(R - \frac{2c}{\mu}\right) \\
&\Longleftrightarrow RV_d(1) > RV_d(2).
\end{aligned}$$

This and the fact that $RV_d(\cdot)$ is unimodal by Lemma 17 imply that when the service fee is set to maximize the revenue, the resulting optimal balking threshold is $k_d^{**} = 1$. As a result, the average sojourn time in the dedicated system is $W_d^{**} = \frac{1}{\mu}$, and the corresponding optimal service fee is $f_d^{**} = R - \frac{c}{\mu}$. Thus, $R - f_d^{**} - cW_d^{**} = 0$, implying that the dedicated system's consumer surplus is 0 and the dedicated system's social welfare with revenue-maximizing fee is

$$SW_d^{**} = RV_d^{**}. \tag{A.248}$$

Using the same arguments as in the proof of Proposition 1, one can show that the throughput of the observable pooled system is larger than that of the observable dedicated system for any fixed fee. Then, the maximum revenues (at the revenue-maximizing fees) satisfy the following for the observable pooled and observable dedicated systems:

$$RV_p^{**} \geq RV_d^{**}. \tag{A.249}$$

This is because with any fixed service fee, the observable pooled system results in larger revenue than the observable dedicated system. Combining (A.248) and (A.249), we have

$$SW_p^{**} \geq RV_p^{**} \geq RV_d^{**} = SW_d^{**} \tag{A.250}$$

when $R/c < \widehat{\eta}$, i.e., $R/c < \left(\frac{(1+\rho)^2}{\rho} + 1\right)\frac{1}{\mu}$. Here, SW_j^{**} represents the social welfare of the observable system $j \in \{p, d\}$ when the fee is set to maximize revenue.

We already know from Lemma 23 that when the fee is set to maximize revenue, the social welfare in the unobservable pooled system is strictly larger than that in the observable pooled system, i.e., $\widehat{SW}_p^{**} > SW_p^{**}$, if $\frac{R\mu}{c} > \frac{N+1}{N}$ and $\lambda < \bar{\lambda}$. Combining this, (A.250) and Proposition 8 - (b), the claim in Lemma 4 follows. Note that $\left(\frac{(1+\rho)^2}{\rho} + 1\right)\frac{1}{\mu} > \frac{3}{\mu} > \frac{N+1}{N}\frac{1}{\mu}$. Thus, the set $\left(\frac{N+1}{N}\frac{1}{\mu}, \left(\frac{(1+\rho)^2}{\rho} + 1\right)\frac{1}{\mu}\right)$ is non-empty. □

APPENDIX B

PROOF OF RESULTS IN CHAPTER 3

B.1 Proof of Lemma 5

It follows from an application of Bayes rule that

$$\begin{aligned}
y_t &= \mathbb{P}(p = p_H | y_{t-1}, X_t) \\
&= \frac{\mathbb{P}(p = p_H, X_t | y_{t-1})}{\mathbb{P}(X_t | y_{t-1})} \\
&= \frac{\mathbb{P}(p = p_H, X_t | y_{t-1})}{\mathbb{P}(p = p_H, X_t | y_{t-1}) + \mathbb{P}(p = p_L, X_t | y_{t-1})} \\
&= \frac{\mathbb{P}(X_t | p = p_H, y_{t-1}) \mathbb{P}(p = p_H | y_{t-1})}{\mathbb{P}(X_t | p = p_H, y_{t-1}) \mathbb{P}(p = p_H | y_{t-1}) + \mathbb{P}(X_t | p = p_L, y_{t-1}) \mathbb{P}(p = p_L | y_{t-1})} \\
&= \frac{p_H^{X_t}(1 - p_H)^{1 - X_t} y_{t-1}}{p_H^{X_t}(1 - p_H)^{1 - X_t} y_{t-1} + p_L^{X_t}(1 - p_L)^{1 - X_t}(1 - y_{t-1})}.
\end{aligned}$$

□

B.2 Proof of Proposition 10

Proof of Part (a): Recall the optimality equation from (3.9),

$$V_F(n, y) = \begin{cases} \max \Big\{ 0, -c + (y p_H + (1 - y) p_L) V_F \left(n - 1, \frac{p_H y}{p_H y + p_L (1 - y)} \right) + \\ \qquad\qquad (y(1 - p_H) + (1 - y)(1 - p_L)) V_F \left(n, \frac{(1 - p_H) y}{(1 - p_H) y + (1 - p_L)(1 - y)} \right) \Big\} & \text{if } n \geq 1, \\ R & \text{if } n = 0. \end{cases}$$

Define

$$p(y) \doteq yp_H + (1-y)p_L, \tag{B.1}$$

$$g_1(y) \doteq \frac{p_H y}{p_H y + p_L(1-y)}, \tag{B.2}$$

$$g_2(y) \doteq \frac{(1-p_H)y}{(1-p_H)y + (1-p_L)(1-y)}, \tag{B.3}$$

$$f(n,y) \doteq \begin{cases} -c + p(y)V_F(n-1, g_1(y)) + (1-p(y))V_F(n, g_2(y)), & \text{if } n \geq 1, \\ R, & \text{if } n = 0. \end{cases} \tag{B.4}$$

Then the optimality equation for dynamic programming can be rewritten as below:

$$V_F(n,y) = \max\{0, f(n,y)\}. \tag{B.5}$$

We will use induction to prove $V_F(n,y)$ is decreasing in n, i.e., $V_F(n,y) \leq V_F(n-1,y)$ for any $n \in \mathbb{N}_+$ and $y \in [0,1]$.

When $n = 1$, $V_F(n,y) \leq V_F(n-1,y)$ since $V_F(1,y) \leq R = V_F(0,y)$ for any $y \in [0,1]$.

Suppose $V_F(n,y) \leq V_F(n-1,y)$ when $n = 1, 2, \ldots, k-1$. Now we need to show that it is true when $n = k$, i.e.,$V_F(k,y) \leq V_F(k-1,y)$ for any $y \in [0,1]$.

Suppose it does not hold for $n = k$, then there exists $y \in [0,1]$ such that $V_F(k,y) > V_F(k-1,y)$. Let

$$d_1 \doteq V_F(k,y) - V_F(k-1,y) > 0. \tag{B.6}$$

By (B.5), we have

$$\tag{B.7}$$

$$V_F(k,y) = \max\{0, f(k,y)\} \tag{B.8}$$

$$V_F(k-1,y) = \max\{0, f(k-1,y)\}. \tag{B.9}$$

Combining (B.6), (B.8) and (B.9), we have

$$f(k,y) - f(k-1,y) \geq d_1 \tag{B.10}$$

since $V_F(k,y) = f(k,y)$ and $V_F(k-1,y) \geq f(k-1,y)$.

Recall the definition of $f(k,y)$ from (B.4). Since $k \geq 2$, we have

$$f(k,y) = -c + p(y)V_F(k-1, g_1(y)) + (1-p(y))V_F(k, g_2(y)) \tag{B.11}$$

$$f(k-1,y) = -c + p(y)V_F(k-2, g_1(y)) + (1-p(y))V_F(k-1, g_2(y)). \tag{B.12}$$

Combining (B.10), (B.11), (B.12), it follows that

$$p(y)(V_F(k-1,g_1(y)) - V_F(k-2,g_1(y))) + (1-p(y))(V_F(k,g_2(y)) - V_F(k-1,g_2(y))) \geq d_1$$

$$\Rightarrow V_F(k,g_2(y)) - V_F(k-1,g_2(y)) \geq d_1/(1-p(y)) \tag{B.13}$$

$$\Rightarrow V_F(k,g_2(y)) - V_F(k-1,g_2(y)) \geq d_1/(1-p_L). \tag{B.14}$$

The inequality (B.13) follows from $V_F(k-1,g_1(y)) \leq V_F(k-2,g_1(y))$ by induction hypothesis. The inequality (B.14) follows from the fact that $1-p_H \leq 1-p(y) \leq 1-p_L$ since $p(y) = yp_H + (1-y)p_L \in [p_L, p_H]$.

Replace y with $g_2(y)$, then $y \in [0,1]$ and $V(k,y) - V(k-1,y) \geq d_1/(1-p_L)$.

Let

$$h_1 \doteq \left\lceil \frac{\ln(d_1/R)}{\ln(1-p_L)} \right\rceil + 1. \tag{B.15}$$

After repeating the above procedure h_1 times, we get $y \in [0,1]$ and $V_F(k,y) - V_F(k-1,y)) \geq d_1/(1-p_L)^{h_1} > R$. It contradicts with the fact that $0 \leq V_F(k,y) \leq R$ for any $y \in [0,1]$. Thus, for any $y \in [0,1]$, $V_F(k,y) \leq V_F(k-1,y)$.

By induction, $V_F(n,y)$ is decreasing in n for fixed y. $\square$

Proof of Part (b): We will use induction to prove this claim.

When $n = 0$, $V_F(n,y)$ is increasing in y since $V_F(0,y) = R$ for any $y \in [0,1]$.

Suppose $V_F(n, y)$ is increasing in y when $n = 0, 1, 2, \ldots, k-1$. We need to prove that $V_F(k, y)$ is also increasing in y.

Suppose there exist $y_1, y_2 \in [0, 1]$ such that $y_1 < y_2$ and $V_F(k, y_1) > V_F(k, y_2)$. Let

$$d_2 \doteq V_F(k, y_1) - V_F(k, y_2) > 0. \tag{B.16}$$

By (B.5), we have

$$V_F(k, y_1) = \max\{0, f(k, y_1)\} \tag{B.17}$$

$$V_F(k, y_2) = \max\{0, f(k, y_2)\}. \tag{B.18}$$

Combing (B.16), (B.17) and (B.18),

$$f(k, y_1) - f(k, y_2) \geq d_2. \tag{B.19}$$

The inequality (B.19) is because $V_F(k, y_1) > V_F(k, y_2) \geq 0$, which implies $V_F(k, y_1) = f(k, y_1)$, and $V_F(k, y_2) \geq f(k, y_2)$.

Recall the definition of $f(n, y)$ from (B.4). Since $k \geq 1$, we have

$$\begin{aligned} f(k, y_1) &= -c + p(y_1)V_F(k-1, g_1(y_1)) + (1 - p(y_1))V_F(k, g_2(y_1)) \\ &\leq -c + p(y_2)V_F(k-1, g_1(y_1)) + (1 - p(y_2))V_F(k, g_2(y_1)) \qquad \text{(B.20)} \\ &\leq -c + p(y_2)V_F(k-1, g_1(y_2)) + (1 - p(y_2))V_F(k, g_2(y_1)) \qquad \text{(B.21)} \\ f(k, y_2) &= -c + p(y_2)V_F(k-1, g_1(y_2)) + (1 - p(y_2))V_F(k, g_2(y_2)). \qquad \text{(B.22)} \end{aligned}$$

The inequality (B.20) is because $p(y_1) < p(y_2)$ when $y_1 < y_2$ and $V_F(k-1, g_1(y)) \geq V_F(k-1, g_2(y)) \geq V_F(k, g_2(y))$, which follows from the induction hypothesis, $g_1(y) \geq y \geq g_2(y)$ and Proposition 10-(a). The inequality (B.21) follows from the induction hypothesis and the fact that $g_1(y_1) < g_1(y_2)$ when $y_1 < y_2$.

Combining (B.19), (B.21) and (B.22), we have

$$V_F(k, g_2(y_1)) - V_F(k, g_2(y_2)) \geq d_2/(1 - p(y_2)) \geq d_2/(1 - p_L). \tag{B.23}$$

The second " $\geq$ " in the inequality (B.23) is because $p(y_2) \geq p_L$ for any $y_2 \in [0, 1]$.

Set $y_1 = g_2(y_1)$ and $y_2 = g_2(y_2)$, then $y_1, y_2 \in [0, 1]$ and $y_1 < y_2$ since $g(y)$ is strictly increasing in y. Then $V_F(k, y_1) - V_F(k, y_2) \geq d_2/(1 - p_L)$ according to (B.23).

Let

$$h_2 \doteq \left\lceil \frac{\ln(d_2/R)}{\ln(1 - p_L)} \right\rceil + 1. \tag{B.24}$$

After repeating the above procedure h_2 times, we get $y_1, y_2 \in [0, 1]$ and $y_1 < y_2$, $V_F(k, y_1)) - V_F(k, y_2)) \geq d_2/(1 - p_L)^{h_2} > R$. It contradicts with the fact that $0 \leq V_F(k, y) \leq R$ for any $y \in [0, 1]$. Thus, $V_F(k, y)$ is also increasing in y.

By induction, $V_F(n, y)$ is increasing in y. $\square$

Proof of Part (c): Consider two situations with different reward R_1 and R_2, $R_1 > R_2$, while other parameters are same. Let π_1^* and π_2^* denote the forward-looking policy for each situation, respectively. And let $V_F^{(1)}(n, y)$ and $V_F^{(2)}(n, y)$ denote the expected total benefit, respectively.

Recall the $V_F(n, y)$ from (3.8),

$$V_F^{(j)}(n, y) \doteq \max \mathbb{E}\left[\sum_{t=0}^{T} r_j(N_t, A_t) | N_0 = n, \alpha_0 = y\right]. \tag{B.25}$$

where $r_j(N_t, A_t)$ is the reward in time period t given the action is A_t for system j:

$$r_j(N_t, A_t) = \begin{cases} -c & \text{if } A_t = 1 \\ R_j & \text{if } A_t = 0 \text{ and } N_t = 0 \\ 0 & \text{if } A_t = 0 \text{ and } N_t > 0 \end{cases} \tag{B.26}$$

Thus,

$$\begin{aligned} V_F^{(1)}(n,y) =& \mathbb{E}_{\pi_1^*}\left[\sum_{t=0}^{T} r_1(N_t,A_t)|N_0=n,\alpha_0=y\right] \\ \geq& \mathbb{E}_{\pi_2^*}\left[\sum_{t=0}^{T} r_1(N_t,A_t)|N_0=n,\alpha_0=y\right] \qquad \text{(B.27)} \\ \geq& \mathbb{E}_{\pi_2^*}\left[\sum_{t=0}^{T} r_2(N_t,A_t)|N_0=n,\alpha_0=y\right] \qquad \text{(B.28)} \\ =& V_F^{(2)}(n,y) \end{aligned}$$

The inequality (B.27) is because π_1^* is the forward-looking policy when the reward is R_1. The inequality (B.28) is because $r_1(N_t, A_t) \geq r_2(N_t, A_t)$ for any (N_t, A_t).

Thus, $V(n, y)$ is increasing in R. □

Proof of Part (d): The proof is similar to part (c).

Consider two situations with different costs c_1 and c_2, $c_1 < c_2$, while other parameters are same. Let $\bar{\pi}_1^*$ and $\bar{\pi}_2^*$ denote the forward-looking policy for each situation, respectively. And let $\bar{V}_F^{(1)}(n, y)$ and $\bar{V}_F^{(2)}(n, y)$ denote the expected total benefit, respectively.

Recall the $V_F(n, y)$ from (3.8),

$$\bar{V}_F^{(j)}(n,y) \doteq \max \mathbb{E}\left[\sum_{t=0}^{T} \bar{r}_j(N_t,A_t)|N_0=n,\alpha_0=y\right]. \qquad \text{(B.29)}$$

where $\bar{r}_j(N_t, A_t)$ is given as below:

$$\bar{r}_j(N_t,A_t) = \begin{cases} -c_j & \text{if } A_t = 1 \\ R & \text{if } A_t = 0 \text{ and } N_t = 0 \\ 0 & \text{if } A_t = 0 \text{ and } N_t > 0 \end{cases} \qquad \text{(B.30)}$$

Thus,

$$\begin{aligned}
\bar{V}_F^{(1)}(n,y) =& \mathbb{E}_{\bar{\pi}_1^*}\left[\sum_{t=0}^{T}\bar{r}_1(N_t,A_t)|N_0=n,\alpha_0=y\right] \\
\geq& \mathbb{E}_{\bar{\pi}_2^*}\left[\sum_{t=0}^{T}\bar{r}_1(N_t,A_t)|N_0=n,\alpha_0=y\right] \qquad \text{(B.31)} \\
\geq& \mathbb{E}_{\bar{\pi}_2^*}\left[\sum_{t=0}^{T}\bar{r}_2(N_t,A_t)|N_0=n,\alpha_0=y\right] \qquad \text{(B.32)} \\
=& \bar{V}_F^{(2)}(n,y)
\end{aligned}$$

The inequality (B.31) is because $\bar{\pi}_1^*$ is the forward-looking policy when the cost per unit time is c_1. The inequality (B.32) is because $\bar{r}_1(N_t,A_t) \geq \bar{r}_2(N_t,A_t)$ for any (N_t,A_t).

Thus, $V(n,y)$ is decreasing in c. □

Proof of Part (e):

Since n is an non-negative integer, in order to prove the convexity of $V_F(n,y)$ in n, it is sufficient to show that $V_F(n+1,y)+V_F(n-1,y) \geq 2V_F(n,y)$ for any $y \in [0,1]$ and $n \in \{1,2,\dots\}$.

First, we prove that $V_F(n+1,y)+V_F(n-1,y) \geq 2V_F(n,y)$ is true for $n=1$ and any $y \in [0,1]$, i.e., $V_F(2,y)+V_F(0,y) \geq 2V_F(1,y)$ for any $y \in [0,1]$.

$V_F(0,y) = R$ for any $y \in [0,1]$. We consider two cases for $V_F(1,y)$.

Case 1: $V_F(1,y) = 0$. Then $V_F(2,y) = 0$ since $V_F(n,y)$ is decreasing in n by Proposition 10-(a). $V_F(2,y)+V_F(0,y) = R \geq 2V_F(1,y)$.

Case 2: $V_F(1,y) > 0$. Recall the optimality equation from (3.9), we have

$$\begin{aligned}
V_F(1,y) =& -c+(yp_H+(1-y)p_L)R+(y(1-p_H)+(1-y)(1-p_L))V_F(1,g_2(y)) \\
\leq& -c+(yp_H+(1-y)p_L)R+(y(1-p_H)+(1-y)(1-p_L))V_F(1,y) \qquad \text{(B.33)} \\
\Rightarrow V_F(1,y) \leq& R-\frac{c}{yp_H+(1-y)p_L} \qquad \text{(B.34)}
\end{aligned}$$

(B.33) follows from the fact that $g_2(y) \leq y$ and $V_F(n,y)$ is increasing in y according to Proposition 10-(b).

Next, we consider two cases for $V_F(2,y)$.

Case 2.1 $V_F(2,y)=0$. Recall the optimality equation from (3.9), we have

$$-c+(yp_H+(1-y)p_L)V_F(1,g_1(y))+(y(1-p_H)+(1-y)(1-p_L))V_F(2,g_2(y))\leq 0 \quad \text{(B.35)}$$

$$\Rightarrow V_F(1,y)\leq V_F(1,g_1(y))\leq \frac{c}{yp_H+(1-y)p_L}. \quad \text{(B.36)}$$

The first " $\leq$ " in (B.36) is because $V_F(n,y)$ is increasing in y according to Proposition 10-(b) and $y\leq g_1(y)$. The second " $\leq$ " in (B.36) follows from (B.35) and the fact that $V_F(2,g_2(y))\geq 0$. Combing (B.34) and (B.36), it implies that

$$2V_F(1,y)\leq R.$$

Thus,

$$V_F(2,y)+V_F(0,y)=R\geq 2V_F(1,y).$$

Case 2.2 $V_F(2,y)>0$.

Consider two systems in parallel. In the first system, there is one forward-looking customer in the queue with state $(1,y)$ while another customer is in service. This forward-looking customer will use the optimal policy. In the second system, there are two customers in the queue waiting for the service. The customer in the first position in queue is a simple customer and he will not leave the system until he completes the service. The customer in the second position in queue use the following policy: before the service completion, if the forward-looking customer in the first system leaves, then she will follow the forward-looking customer to leave. If she does not leave before the customer in service finishes the service, then she moves to the first position in queue and apply the optimal policy for state $(1,y)$ from this point.

Let τ_1 denotes the random variable for the time till the customer in service leaves the system for the customer in state $(1,y)$, and let τ_2 denotes the random variable for the time till the customer in state $(1,y)$ abandons the service. If the customer doesn't abandon the service, $\tau_2=+\infty$.

$$V_F(1,y)=\mathbb{E}(R-c\tau_1|\tau_1\leq\tau_2)\mathbb{P}(\tau_1\leq\tau_2)-c\mathbb{E}(\tau_2|\tau_1>\tau_2)(1-\mathbb{P}(\tau_1\leq\tau_2)) \quad \text{(B.37)}$$

The equation (B.37) is because the customer will get an expected benefit of $R - c\tau_1$ if he does not leave until she is in service at τ_1, and the expected total benefit is $-c\tau_2$ is she leaves the system before getting served at τ_2. Suppose the customer with state $(2, y)$ uses the same policy for state $(1, y)$ until a service completes. Thus,

$$V_F(2,y) \geq \mathbb{E}(V_F(1,y) - c\tau_1 | \tau_1 \leq \tau_2)\mathbb{P}(\tau_1 \leq \tau_2) - c\mathbb{E}(\tau_2 | \tau_1 > \tau_2)(1 - \mathbb{P}(\tau_1 \leq \tau_2)). \tag{B.38}$$

The inequality (B.38) follows from the fact that $V_F(2, y)$ is the expected value the customer in state $(2, y)$ gets with the optimal policy and $\mathbb{E}(V_F(1, y) - c\tau_1 | \tau_1 \leq \tau_2)\mathbb{P}(\tau_1 \leq \tau_2) - c\mathbb{E}(\tau_2 | \tau_1 > \tau_2)(1 - \mathbb{P}(\tau_1 \leq \tau_2))$ is the value she can get with the policy described. By (B.37) and (B.38), we have

$$\begin{aligned} & V_F(1,y) - V_F(2,y) \leq (R - V_F(1,y))\mathbb{P}(\tau_1 \leq \tau_2) \leq R - V_F(1,y) \\ \Rightarrow & 2V_F(1,y) \leq V_F(0,y) + V_F(2,y) \end{aligned} \tag{B.39}$$

Thus, $V_F(0, y) + V_F(2, y) \geq 2V_F(1, y)$.

Suppose $V_F(n + 1, y) + V_F(n - 1, y) \geq 2V_F(n, y)$ is true for any y when $n = 1, 2, \ldots, k - 1$.

When $n = k$, if there exists $y \in [0, 1]$ such that $V_F(n + 1, y) + V_F(n - 1, y) < 2V_F(n, y)$. Let

$$d_3 \doteq 2V_F(k,y) - V_F(k+1,y) - V_F(k-1,y) > 0. \tag{B.40}$$

Then $V_F(k, y) > 0$ and $V_F(k - 1, y) \geq V_F(k, y) > 0$ since $V_F(n, y)$ is decreasing in n by Proposition 10-(a).

Recall the optimality equation from (3.9). Since $k \geq 2$, we have

$$V_F(k,y) = -c + (yp_H + (1-y)p_L)V_F(k-1, g_1(y)) + (y(1-p_H) + (1-y)(1-p_L))V_F(k, g_2(y)) \tag{B.41}$$

$$\begin{aligned} V_F(k-1,y) = & -c + (yp_H + (1-y)p_L)V_F(k-2, g_1(y)) \\ & + (y(1-p_H) + (1-y)(1-p_L))V_F(k-1, g_2(y)) \end{aligned} \tag{B.42}$$

$$V_F(k+1,y) \geq -c + (yp_H + (1-y)p_L)V_F(k, g_1(y)) + (y(1-p_H) + (1-y)(1-p_L))V_F(k+1, g_2(y)) \tag{B.43}$$

By (B.41), (B.42) and (B.43), we have

$$
\begin{aligned}
&(yp_H + (1-y)p_L)(2V_F(k-1, g_1(y)) - V_F(k-2, g_1(y)) - V_F(k, g_1(y))) \\
&+(y(1-p_H) + (1-y)(1-p_L))(2V_F(k, g_2(y)) - V_F(k-1, g_2(y)) - V_F(k+1, g_2(y))) \\
&\geq 2V_F(k, y) - V_F(k-1, y) - V_F(k+1, y) = d_3 \\
&\Rightarrow (y(1-p_H) + (1-y)(1-p_L))(2V_F(k, g_2(y)) - V_F(k-1, g_2(y)) - V_F(k+1, g_2(y))) \geq d_3 \quad \text{(B.44)} \\
&\Rightarrow 2V_F(k, g_2(y)) - V_F(k-1, g_2(y)) - V_F(k+1, g_2(y)) \geq \frac{d_3}{1-p_L} \quad \text{(B.45)}
\end{aligned}
$$

The inequality (B.44) follows from the fact that $2V_F(k-1, g_1(y)) - V_F(k-2, g_1(y)) - V_F(k, g_1(y)) \leq 0$ by induction hypothesis. The inequality (B.45) follows from the fact that $1 - p_H \leq (y(1-p_H) + (1-y)(1-p_L)) \leq 1 - p_L$.

Replace y with $g_2(y)$, we get $y \in [0,1]$ and $2V_F(k, y) - V_F(k-1, y) - V_F(k+1, y) \geq \frac{d_3}{1-p_L}$.

Let

$$h_3 \doteq \left\lceil \frac{\ln(d_3/R)}{\ln(1-p_L)} \right\rceil + 1. \quad \text{(B.46)}$$

After repeating the above procedure h_3 times, we gat $y \in [0,1]$ and

$$2V_F(k, g_2(y)) - V_F(k-1, g_2(y)) - V_F(k+1, g_2(y)) \geq \frac{d_3}{(1-p_L)^{h_3}} > R. \quad \text{(B.47)}$$

It contradicts with the fact that $2V_F(k, g_2(y)) - V_F(k-1, g_2(y)) - V_F(k+1, g_2(y)) \leq V_F(k, g_2(y)) \leq R$.

Thus, $V_F(n+1, y) + V_F(n-1, y) \geq 2V_F(n, y)$ for any y when $n = k$.

By induction, $V_F(n+1, y) + V_F(n-1, y) \geq 2V_F(n, y)$ is true for any non-negative integer n and any $y \in [0,1]$. Thus, $V_F(n, y)$ is convex in n. □

B.3 Proof of Proposition 11

Proof of Part (a): We prove two lemmas first, which will be used in the remainder of the proof.

Lemma 24. *$V_F(n, 0) = \max\{R - \frac{cn}{p_L}, 0\}$ and $V_F(n, 1) = \max\{R - \frac{cn}{p_H}, 0\}$.*

Proof of Lemma 24: Recall the optimality equation from (3.9),

$$\begin{aligned}
V_F(n,0) =& \max\{0, -c + p_L V_F(n-1,0) + (1-p_L)V_F(n,0)\},\\
V_F(n,1) =& \max\{0, -c + p_H V_F(n-1,1) + (1-p_H)V_F(n,1)\},\\
V_F(0,0) =& R,\\
V_F(0,1) =& R.
\end{aligned}$$

We consider two cases for $V_F(n,0)$. Case 1: If $V_F(n-1,0) \geq c/p_L$, then $-c + p_L V_F(n-1,0) + (1-p_L)V_F(n,0) \geq 0$ and then $V_F(n,0) = -c + p_L V_F(n-1,0) + (1-p_L)V_F(n,0)$, which implies $V_F(n,0) = V_F(n-1,0) - c/p_L$. Case 2: If $V_F(n-1,0) < c/p_L$, then $-c + p_L V_F(n-1,0) + (1-p_L)V_F(n,0) < (1-p_L)V_F(n,0)$. Suppose $V_F(n,0) > 0$, then $V_F(n,0) < (1-p_L)V_F(n,0)$. It contradicts with $V_F(n,0) > 0$ and thus $V_F(n,0) = 0$.

Based on these and the fact that $V_F(0,0) = R$,

$$V_F(n,0) = \max\left\{0, R - \frac{cn}{p_L}\right\}. \tag{B.48}$$

Similarly, we can find $V_F(n,1)$ by replacing p_L with p_H in the above analysis. We have

$$V_F(n,1) = \max\left\{0, R - \frac{cn}{p_H}\right\}. \tag{B.49}$$

□

Lemma 25. *$f(n,y)$ defined in (B.4) is increasing in y for any given n.*

Proof of Lemma 25: When $n = 0$, $f(n,y) = R$ and thus $f(n,y)$ is increasing in y.

When $n \geq 1$, for $y_1, y_2 \in [0,1]$ and $y_1 < y_2$, it is easy to show that $p(y_1) < p(y_2)$, $g_1(y_1) < g_1(y_2)$ and $g_2(y_1) < g_2(y_2)$.

$$\begin{aligned} f(n,y_1) &= -c + p(y_1)V_F(n-1, g_1(y_1)) + (1-p(y_1))V_F(n, g_2(y_1)) \\ &\leq -c + p(y_2)V_F(n-1, g_1(y_1)) + (1-p(y_2))V_F(n, g_2(y_1)) \qquad \text{(B.50)} \\ &\leq -c + p(y_2)V_F(n-1, g_1(y_2)) + (1-p(y_2))V_F(n, g_2(y_2)) \qquad \text{(B.51)} \\ &= f(n,y_2) \end{aligned}$$

The inequality (B.50) follows from the fact that $p(y_1) < p(y_2)$, $V_F(n-1, g_1(y_1)) \geq V_F(n-1, g_2(y_1)) \geq V_F(n, g_2(y_1))$ since $g_1(y_1) \geq g_2(y_1)$ according to Proposition 10-(a)(b), and $p(y_1) < p(y_2)$. The inequality (B.51) follows from $g_1(y_1) < g_1(y_2)$, $g_2(y_1) < g_2(y_2)$ and the fact that $V_F(n,y)$ is increasing in y according to Proposition 10-(b).

Thus, $f(n,y)$ is increasing in y. $\square$

Recall the optimality equation from (B.5),

$$V_F(n,y) = \max\{0, f(n,y)\}.$$

The forward-looking customer will leave if and only if $f(n,y) < 0$.

By Lemma 24, $f(n,1) = R - \frac{cn}{p_H}$. When $n > \frac{Rp_H}{c}$, $f(n,1) = R - \frac{cn}{p_H} < 0$. Thus $f(n,y) < 0$ for $\forall y \in [0,1]$ since $f(n,y)$ is increasing in y by Lemma 25. The customer leaves regardless of y.

For $n \in \{0, 1, 2, \ldots, \lfloor \frac{Rp_H}{c} \rfloor\}$, define

$$\beta_F(n) \doteq \inf\{y : y \in [0,1], f(n,y) \geq 0\}. \qquad \text{(B.52)}$$

Since $f(n,y)$ is increasing in y by Lemma 25, the forward-looking customer will leave whenever $y < \beta_F(n)$, and continue whenever $y \geq \beta_F(n)$. $\square$

Proof of Part (b): By (3.12) and (3.16), the myopic learner or naive customer continues if and only if

$$R - \left(\frac{n}{p_H}y + \frac{n}{p_L}(1-y)\right)c \geq 0 \leftrightarrows y \geq \left(\frac{n}{p_L} - \frac{R}{c}\right) \Big/ \left(\frac{n}{p_L} - \frac{n}{p_H}\right). \qquad \text{(B.53)}$$

Define

$$s(n,y) \doteq R - \left(\frac{n}{p_H}y + \frac{n}{p_L}(1-y)\right)c. \tag{B.54}$$

When $n > \frac{Rp_H}{c}$, $s(n,y) < 0$ for $\forall y \in [0,1]$ and the customer leaves regardless of y.

For $n \in \{0,1,2,\ldots,\lfloor\frac{Rp_H}{c}\rfloor\}$, define

$$\beta_M(n) = \beta_N(n) \doteq \inf\{y : y \in [0,1], s(n,y) \geq 0\}. \tag{B.55}$$

By (B.53), we have

$$\beta_M(n) = \beta_N(n) = \max\left\{\left(\frac{n}{p_L} - \frac{R}{c}\right) \Big/ \left(\frac{n}{p_L} - \frac{n}{p_H}\right), 0\right\}. \tag{B.56}$$

The claim follows. □

B.4 Proof of Proposition 12

Recall the expected total benefit for the forward-looking customer and the naive customer from (B.5) and (3.15), respectively.

$$\begin{aligned} V_F(n,y) &= \max\{0, f(n,y)\}, \\ V_N(n,y) &= \max\left\{0, R - \left(\frac{n}{p_H}y + \frac{n}{p_L}(1-y)\right)c\right\}. \end{aligned}$$

Recall $f(n,y)$ from (B.4).

$$f(n,y) \doteq \begin{cases} -c + p(y)V_F(n-1, g_1(y)) + (1-p(y))V_F(n, g_2(y)), & \text{if } n \geq 1, \\ R, & \text{if } n = 0. \end{cases}$$

Recall $s(n,y)$ from (B.54),

$$s(n,y) = R - \left(\frac{n}{p_H}y + \frac{n}{p_L}(1-y)\right)c. \tag{B.57}$$

Then $V_N(n,y) = \max\{0, s(n,y\}$.

When $n = 0$, $f(n,y) = s(n,y) = R$. When $n \geq 1$, since $p(y) = yp_H + (1-y)p_L$, $g_1(y) = \frac{p_H y}{p_H y + p_L(1-y)}$ and $g_2(y) = \frac{(1-p_H)y}{(1-p_H)y+(1-p_L)(1-y)}$ by (B.2) through (B.3), by elementary calculations we have

$$\begin{aligned} s(n,y) =& -c + p(y)s\left(n-1, g_1(y)\right) + (1-p(y))s\left(n, g_2(y)\right) \\ \leq& -c + p(y)V_N(n-1, g_1(y)) + (1-p(y))V_N(n, g_2(y)) \qquad \text{(B.58)} \\ \leq& -c + p(y)V_F(n-1, g_1(y)) + (1-p(y))V_F(n, g_2(y)) = f(n,y). \qquad \text{(B.59)} \end{aligned}$$

The inequality (B.58) is because $V_N(n,y) = \max(0, s(n,y)) \geq s(n,y)$ for $\forall(n,y)$. The inequality (B.59) follows from the fact that $V_F(n,y) \geq V_N(n,y)$ for $\forall(n,y)$ by the definition of $V_F(n,y)$ from (3.8).

$$\beta_F(n) = \inf\{y : y \in [0,1], f(n,y) \geq 0\}.$$

When the customer with naive policy continues, $s(n,y) \geq 0$.

$$f(n,y) \geq s(n,y) \geq 0.$$

It implies that the customer with forward-looking policy also joins the system.

Thus, $\beta_M(n) = \beta_N(n) \geq \beta_F(n)$ by the definition of $\beta_F(n)$, $\beta_N(n)$ and $\beta_M(n)$ from (B.52) and (B.55).

□

B.5 Proof of Proposition 13

Proof of Part (a): By Lemma 24,

$$\begin{aligned} V_F(n,0) &= \max\left\{0, R - \frac{cn}{p_L}\right\}, \\ V_F(n,1) &= \max\left\{0, R - \frac{cn}{p_H}\right\}. \end{aligned}$$

It implies that $f(n,0) = R - \frac{cn}{p_L}$ and $f(n,1) = R - \frac{cn}{p_H}$.

We consider three cases for the threshold analysis. Case 1: When $n \leq \frac{Rp_L}{c}$, $f(n,0) = R - \frac{cn}{p_L} \geq 0$. Thus, $f(n,y) \geq 0$ for any $y \in [0,1]$ since $f(n,y)$ is increasing in y according to the proof of Proposition 11. It implies that $\beta_F(n) = 0$ by the definition of $\beta_F(n)$ from (B.52). Case 2: When $n > \frac{Rp_L}{c}$ and $n < \frac{Rp_H}{c}$, $f(n,0) = R - \frac{cn}{p_L} < 0$ and $f(n,1) = R - \frac{cn}{p_H} > 0$, which implies $\beta_F(n) \in (0,1)$. Case 3: When $n = \frac{Rp_H}{c}$, $f(n,1) = R - \frac{cn}{p_H} = 0$ and $f(n,y) < R - \frac{cn}{p_H} = 0$ for any $y \in [0,1)$. It implies $\beta_F(n) = 1$ by the definition of $\beta_F(n)$ from (B.52).

Thus, the threshold for forward-looking policy $\beta_F(n) \in (0,1)$ if and only if $\bar{n}_1 < n < \bar{n}_2$, where $\bar{n}_1 = \frac{Rp_L}{c}$ and $\bar{n}_2 = \frac{Rp_H}{c}$. □

Proof of Part (b): We now show $f(n,y)$ is decreasing in n for given y. Recall $f(n,y)$ from (B.4).

$$f(n,y) \doteq \begin{cases} -c + p(y)V_F(n-1, g_1(y)) + (1-p(y))V_F(n, g_2(y)), & \text{if } n \geq 1, \\ R, & \text{if } n = 0. \end{cases}$$

It is obvious that $f(n,y)$ is decreasing in n since $V_F(n-1, g_1(y))$ and $V_F(n, g_2(y))$ are decreasing in n by Proposition 10-(a).

Recall the definition of $\beta_F(n)$ from (B.52).

$$\beta_F(n) = \inf\{y : y \in [0,1], f(n,y) \geq 0\}.$$

Since $f(n,y)$ is increasing in y by the Proof of Proposition 11 and decreasing in n, $\beta_F(n)$ is increasing in n. □

Proof of Part (c): Recall $f(n,y)$ from (B.4),

$$f(n,y) = -c + p(y)V_F(n-1, g_1(y)) + (1-p(y))V_F(n, g_2(y)).$$

By Proposition 10 -(c), $V_F(n,y)$ increases in R and decreases in c. It immediately follows that $f(n,y)$ increases in R and decreases in c since $p(y)$ does not depend on R or c. Thus, the belief threshold $\beta_F(n) \doteq \inf\{y : y \in [0,1], f(n,y) \geq 0\}$ decreases in R and increases in c. □

Proof of Part (d): By the threshold result for the myopic learner and the naive customer from (B.56), we have

$$\beta_M(n) = \beta_N(n) = \max\left\{\left(\frac{n}{p_L} - \frac{R}{c}\right) \Big/ \left(\frac{n}{p_L} - \frac{n}{p_H}\right), 0\right\} = \max\left\{\left(\frac{1}{p_L} - \frac{R}{cn}\right) \Big/ \left(\frac{1}{p_L} - \frac{1}{p_H}\right), 0\right\}.$$

Thus, $\beta_M(n) \in (0,1)$ or $\beta_N(n) \in (0,1)$ when $n \in (\bar{n}_1, \bar{n}_2)$, where $\bar{n}_1 = \frac{Rp_L}{c}$, $\bar{n}_2 = \frac{Rp_H}{c}$ as defined in (3.17).

It is obvious that $\beta_M(n)$ and $\beta_N(n)$ increase in n, decrease in R and increase in c. □

B.6 Proof of Proposition 14

When $n = 0$, $f(n,y) = s(n,y) = R$, both types of customers join the system. When $n > \frac{Rp_H}{c}$, both types of customers will not join the system regardless of y by the proof of Proposition 11

We also know that $f(n,y)$ is decreasing in n by the proof of Proposition 13-(b), and it is obvious that $s(n,y) = R - (y\frac{n}{p_H} + (1-y)\frac{n}{p_L})c$ is decreasing in n.

Thus, for any given y, there exist thresholds $n_S(y)$ and $n_F(y)$ such that the naive customer joins the system if and only if $n \leq n_S(y)$, while the forward-looking customer joins the system if and only if $n \leq n_F(y)$.

By Proposition 12, the naive policy results in larger belief threshold for any given n, i.e., $\beta_F(n) \geq \beta_N(n)$. It implies that forward-looking customer will also join when the naive customer joins. Thus, $n_S(y) \leq n_F(y)$. In summary, both forward-looking and naive customers will join for $n \leq n_S(y)$; forward-looking customer joins and naive customer does not join when n in $(n_S(y), n_F(y)]$; both types of customers will not join when $n > n_F(y)$. □

B.7 Proof of Proposition 15

Recall $\bar{n}_1$ from (3.17), $\bar{n}_1 = \frac{Rp_L}{c}$. For any $n \leq \lfloor \bar{n}_1 \rfloor$, $s(n,y) = R - \left(\frac{n}{p_H}y + \frac{n}{p_L}(1-y)\right)c \geq 0$ for $\forall y \in [0,1]$ and thus $\beta_N(n) = 0$ for any $n \leq \lfloor \bar{n}_1 \rfloor$. And $\beta_N(\lfloor \bar{n}_1 \rfloor + 1) \in (0,1)$ since $s(\lfloor \bar{n}_1 \rfloor + 1, 0) = R - \left(\frac{\lfloor \bar{n}_1 \rfloor + 1}{p_L}\right)c < 0$ and $s(\lfloor \bar{n}_1 \rfloor + 1, 1) = R - \left(\frac{\lfloor \bar{n}_1 \rfloor + 1}{p_H}\right)c = \frac{c}{p_H}\left(\frac{Rp_H}{c} - \lfloor \bar{n}_1 \rfloor - 1\right) \geq \frac{c}{p_H}\left(\frac{Rp_H}{c} - \frac{Rp_L}{c} - 1\right) > 0$ when $\frac{R}{c} > \frac{1}{p_H - p_L}$. Thus, $\beta_F(\lfloor \bar{n}_1 \rfloor + 1) < \beta_N(\lfloor \bar{n}_1 \rfloor + 1)$ by Lemma 29.

Let $\underline{y} = \beta_F(\lfloor \bar{n}_1 \rfloor + 1)$ and $\bar{y} = \beta_N(\lfloor \bar{n}_1 \rfloor + 1)$. We already know that $\bar{y} \in (0,1)$, then $\underline{y} < \bar{y} < 1$. In addition, $f(\lfloor \bar{n}_1 \rfloor + 1, 0) = R - c\frac{\lfloor \bar{n}_1 \rfloor + 1}{p_L} < 0$ by Lemma 24 and then $\underline{y} > 0$. Thus, $\underline{y} \in (0,1)$.

When $\alpha_0 \in (\underline{y}, \bar{y})$, the naive customer joins if and only if the number of customers ahead n_0 satisfies $n_0 \leq \lfloor \bar{n}_1 \rfloor$, while the forward-looking customers still join when the number of customers ahead is $\lfloor \bar{n}_1 \rfloor + 1$.

Let $\{X_t^-(\omega), t \in \mathbb{N}^+\}$ and $\{X_t^+(\omega), t \in \mathbb{N}^+\}$ be the number of customers in the system with naive customers at the beginning of each time period before arrival and after arrival, respectively, for a given sample path ω. Let $\{Y_t^-(\omega), t \in \mathbb{N}^+\}$ and $\{Y_t^+(\omega), t \in \mathbb{N}^+\}$ be the number of customers in the system with forward-looking customers at the beginning of each time period before arrival and after arrival, respectively, for a given sample path ω.

We use $\{A_1(\omega), A_2(\omega), \dots\}$ to denote the arrival process and use $\{C_1(\omega), C_2(\omega), \dots\}$ to denote the service completion process in both systems, where

$$A_t(\omega) = \begin{cases} 1 & \text{if there is an arrival at time period t,} \\ 0 & \text{else,} \end{cases} \tag{B.60}$$

$$C_t(\omega) = \begin{cases} 1 & \text{a service is completed at time period t if the system is not empty,} \\ 0 & \text{else.} \end{cases} \tag{B.61}$$

Each system is empty at the beginning, i.e., $X_1^-(\omega) = Y_1^-(\omega) = 0$ and then $X_1^+(\omega) = Y_1^+(\omega) = A_1(\omega)$.

Suppose $Y_t^-(\omega) \geq X_t^-(\omega)$ when $t = 1, \dots, i$. We will consider two cases and prove that $Y_i^+(\omega) \geq X_i^+(\omega)$ and $Y_{i+1}^-(\omega) \geq X_{i+1}^-(\omega)$.

Case 1: $Y_i^-(\omega) \leq \lfloor \bar{n}_1 \rfloor$. Then $X_i^-(\omega) \leq Y_i^-(\omega) \leq \lfloor \bar{n}_1 \rfloor$ by induction hypothesis. We consider four sub-cases as below: Case 1.1: If $A_i(\omega) = 1$ and $C_i(\omega) = 0$, since $\beta_N(n_i) \leq 0$ and $\beta_F(n_i) \leq 0$ for any $n_i \leq \lfloor \bar{n}_1 \rfloor$, the arrival with initial belief $\alpha_0 \in (\underline{y}, \bar{y})$ will join for both systems. Thus, $X_i^+(\omega) = X_i^-(\omega) + 1 \leq Y_i^-(\omega) + 1 = Y_i^+(\omega) \leq \lfloor \bar{n}_1 \rfloor + 1$. Since $\beta_N(n_i) \leq 0$ and $\beta_F(n_i) \leq 0$ for any $n_i \leq \lfloor \bar{n}_1 \rfloor$, no strategic customer will abandon at the end of time period i. Then $X_{i+1}^-(\omega) = X_i^+(\omega) = X_i^-(\omega) + 1$ and $Y_{i+1}^-(\omega) = Y_i^+(\omega) = Y_i^-(\omega) + 1$. Thus, $X_{i+1}^-(\omega) \leq Y_{i+1}^-(\omega)$. Case 1.2: If $A_i = 0$ and $C_i = 0$, $X_{i+1}^-(\omega) = X_i^+(\omega) = X_i^-(\omega)$ and $Y_{i+1}^-(\omega) = Y_i^+(\omega) = Y_i^-(\omega)$ since $\beta_F(n_i) \leq 0$ for any $n_i \leq \lfloor \bar{n}_1 \rfloor$ and then no forward-looking customer abandons the service. We also have that $X_i^+(\omega) \leq Y_i^+(\omega)$ and $X_{i+1}^-(\omega) \leq Y_{i+1}^-(\omega)$. Case 1.3: If $A_i(\omega) = 1$ and $C_i(\omega) = 1$, $X_i^+(\omega) = X_i^-(\omega) + 1$ and $Y_i^+(\omega) = Y_i^-(\omega) + 1$. Thus $X_i^+(\omega) \leq Y_i^+(\omega)$. And $X_{i+1}^-(\omega) \leq Y_{i+1}^-(\omega)$ since $X_{i+1}^-(\omega) = X_i^-(\omega)$ and $Y_{i+1}^-(\omega) = Y_i^-(\omega)$. Case 1.4:

If $A_i(\omega) = 0$ and $C_i(\omega) = 1$, $X_i^+(\omega) = X_i^-(\omega) \leq Y_i^-(\omega) = Y_i^+(\omega)$. And $X_{i+1}^-(\omega) \leq Y_{i+1}^-(\omega)$ since $X_{i+1}^-(\omega) = \max\{0, X_i^+(\omega) - 1\}$ and $Y_{i+1}^-(\omega) = \max\{0, Y_i^+(\omega) - 1\}$.

Case 2: $Y_i^-(\omega) \geq \lfloor \bar{n}_1 \rfloor + 1$. We know that $X_i^-(\omega) \leq \lfloor \bar{n}_1 \rfloor + 1$ and $X_i^+(\omega) \leq \lfloor \bar{n}_1 \rfloor + 1$ since $\alpha_0 < \beta_N(\lfloor \bar{n}_1 \rfloor + 1)$. Case 2.1: If $A_i(\omega) = 1$ and $C_i(\omega) = 0$, $Y_i^+(\omega) \geq \lfloor \bar{n}_1 \rfloor + 1 \geq X_i^+(\omega)$, and $Y_{i+1}^-(\omega) \geq \lfloor \bar{n}_1 \rfloor + 1 \geq X_{i+1}^-(\omega)$ since $\beta_F(n_i) \leq 0$ for any $n_i \leq \lfloor \bar{n}_1 \rfloor$. Case 2.2: If $A_i = 0$ and $C_i = 0$, $Y_{i+1}^-(\omega) \geq \lfloor \bar{n}_1 \rfloor + 1 \geq X_{i+1}^-(\omega)$ since $\beta_F(n_i) \leq 0$ for any $n_i \leq \lfloor \bar{n}_1 \rfloor$. Case 2.3: If $A_i(\omega) = 1$ and $C_i(\omega) = 1$, $Y_i^+(\omega) \geq \lfloor \bar{n}_1 \rfloor + 1 \geq X_i^+(\omega)$, and $Y_{i+1}^-(\omega) \geq \lfloor \bar{n}_1 \rfloor + 1 \geq X_{i+1}^-(\omega)$. Case 2.4: If $A_i = 0$ and $C_i = 1$, $Y_i^+(\omega) \geq \lfloor \bar{n}_1 \rfloor + 1 \geq X_i^+(\omega)$, and $Y_{i+1}^-(\omega) \geq \lfloor \bar{n}_1 \rfloor \geq X_{i+1}^-(\omega)$.

Thus, $X_i^+(\omega) \leq Y_i^+(\omega)$ and $X_{i+1}^-(\omega) \leq Y_{i+1}^-(\omega)$. By induction, $X_t^-(\omega) \leq Y_t^-(\omega)$ and $X_t^+(\omega) \leq Y_t^+(\omega)$ for any t.

The average throughput $\theta_F(\tau, \omega)$ and $\theta_N(\tau, \omega)$ for finite time horizon τ are defined as below:

$$\theta_N(\tau, \omega) = \sum_{t=1}^{\tau} C_t(\omega) I_{\{X_t^+(\omega)>0\}}/\tau \tag{B.62}$$

$$\theta_F(\tau, \omega) = \sum_{t=1}^{\tau} C_t(\omega) I_{\{Y_t^+(\omega)>0\}}/\tau \tag{B.63}$$

Thus,

$$\theta_N(\tau, \omega) = \sum_{t=1}^{\tau} C_t(\omega) I_{\{X_t^+(\omega)>0\}}/\tau \leq \sum_{t=1}^{\tau} C_t(\omega) I_{\{Y_t^+(\omega)>0\}}/\tau = \theta_F(\tau, \omega)$$

Moreover, it is true for any given sample path ω based on the above analysis.

We construct a set of sample paths Ω_1 as below. $\omega \in \Omega_1$ if and only if $A_t(\omega) = 1$ and $C_t(\omega) = 0$ for $t \in \{1, 2, \ldots, \lfloor \bar{n}_1 \rfloor + 1\}$, $A_t(\omega) = 1$ and $C_t(\omega) = 1$ for $t = \lfloor \bar{n}_1 \rfloor + 2$, $A_t(\omega) = 0$ and $C_t(\omega) = 1$ for $t \in \{\lfloor \bar{n}_1 \rfloor + 3\lfloor \bar{n}_1 \rfloor + 4, \ldots, 2\lfloor \bar{n}_1 \rfloor + 3\}$. Then $X_t^+(\omega) = Y_t^+(\omega) = t$ for $t \in \{1, 2, \ldots, \lfloor \bar{n}_1 \rfloor + 1\}$ since all arrivals join for both systems because $\beta_N(n) \leq 0$ and $\beta_F(n) \leq 0$ when $n \leq \lfloor \bar{n}_1 \rfloor$. $X_{\lfloor \bar{n}_1 \rfloor + 2}^+(\omega) = \lfloor \bar{n}_1 \rfloor + 1$ while $Y_{\lfloor \bar{n}_1 \rfloor + 2}^+(\omega) = \lfloor \bar{n}_1 \rfloor + 2$ since $\alpha_0 \in (\beta_F(\lfloor \bar{n}_1 \rfloor + 1), \beta_N(\lfloor \bar{n}_1 \rfloor + 1))$. And $X_t^+(\omega) = 2\lfloor \bar{n}_1 \rfloor + 3 - t$ while $Y_t^+(\omega) = 2\lfloor \bar{n}_1 \rfloor + 4 - t$ for $t \in \{\lfloor \bar{n}_1 \rfloor + 3, \lfloor \bar{n}_1 \rfloor + 4, \ldots, 2\lfloor \bar{n}_1 \rfloor + 3\}$. We have $I_{\{Y_t^+(\omega)>0\}} \geq I_{\{X_t^+(\omega)>0\}}$ for any t since $Y_t^+(\omega) \geq X_t^+(\omega)$. In addition, $C_t(\omega) I_{\{X_t^+(\omega)>0\}} = 0$ and $C_t(\omega) I_{\{Y_t^+(\omega)>0\}} = 1$ when

$t = 2\lfloor \bar{n}_1 \rfloor + 3$. Thus, when $\tau \geq 2\lfloor \bar{n}_1 \rfloor + 3$, for any $\omega \in \Omega_1$, we have

$$\theta_F(\tau, \omega) = \sum_{t=1}^{\tau} C_t(\omega) I_{\{Y_t^+(\omega)>0\}}/\tau > \sum_{t=1}^{\tau} C_t(\omega) I_{\{X_t^+(\omega)>0\}}/\tau = \theta_N(\tau, \omega)$$

In addition, $\mathbb{P}(\omega \in \Omega_1) > 0$.

Thus, $\theta_F(\tau) = \mathbb{E}_\omega \theta_F(\tau, \omega) > \mathbb{E}_\omega \theta_N(\tau, \omega) = \theta_N(\tau)$ when $\tau \geq 2\lfloor \bar{n}_1 \rfloor + 3$. □

B.8 Proof of Proposition 16

Proof of Part (a): Let

$$\bar{\bar{y}} \doteq \left(\frac{1}{p_L} - \frac{R}{c\bar{n}_2}\right) / \left(\frac{1}{p_L} - \frac{1}{p_H}\right). \tag{B.64}$$

When $\alpha_0 > \bar{\bar{y}}$,

$$\begin{aligned}
\alpha_0 \geq \left(\frac{1}{p_L} - \frac{R}{c\bar{n}_2}\right) / \left(\frac{1}{p_L} - \frac{1}{p_H}\right) &\iff \alpha_0 \left(\frac{1}{p_L} - \frac{1}{p_H}\right) \geq \frac{1}{p_L} - \frac{R}{c\bar{n}_2} \\
&\iff \frac{R}{c\bar{n}_2} \geq \frac{\alpha_0}{p_H} + \frac{1-\alpha_0}{p_L} \\
&\iff R - c\bar{n}_2 \left(\frac{\alpha_0}{p_H} + \frac{1-\alpha_0}{p_L}\right) \geq 0.
\end{aligned}$$

Thus, naive customers will join the system when the number of customers ahead satisfies $n \leq \lfloor \bar{n}_2 \rfloor$ by (3.16). Also, forward-looking customers will join the system when the number of customers ahead satisfies $n \leq \lfloor \bar{n}_2 \rfloor$ since $\beta_F(n) \leq \beta_N(n)$ for any n by Proposition 12. When $n \geq \lfloor \bar{n}_2 \rfloor + 1$, $n > \frac{Rp_H}{c}$ and then the customers of both types will not join by the proof of Proposition 11.

Let $\{X_t^-(\omega), t \in \mathbb{N}^+\}$ and $\{X_t^+(\omega), t \in \mathbb{N}^+\}$ be the number of customers in the system with naive customers at the beginning of each time period before arrival and after arrival, respectively, for a given sample path ω. Let $\{Y_t^-(\omega), t \in \mathbb{N}^+\}$ and $\{Y_t^+(\omega), t \in \mathbb{N}^+\}$ be the number of customers in the system with forward-looking customers at the beginning of each time period before arrival and after arrival, respectively, for a given sample path ω.

Next, we will show that $X_t^-(\omega) \geq Y_t^-(\omega)$ and $X_t^+(\omega) \geq Y_t^+(\omega)$ for any given sample path ω, $t = 1, 2 \ldots$.

$X_1^-(\omega) = Y_1^-(\omega) = 0$, and $X_1^+(\omega) = Y_1^+(\omega) = A_1(\omega)$.

Suppose $X_t^-(\omega) \geq Y_t^-(\omega)$ for $t = 1, \ldots, i$. We will consider two cases below prove that $X_i^+(\omega) \geq Y_i^+(\omega)$ and $X_{i+1}^-(\omega) \geq Y_{i+1}^-(\omega)$.

Case 1: $Y_i^-(\omega) \leq X_i^-(\omega) \leq \lfloor \bar{n}_2 \rfloor$. Case 1.1: If $A_i(\omega) = 1$ and $C_i(\omega) = 0$, $X_i^+(\omega) \geq Y_i^+(\omega)$ since $X_i^+(\omega) = X_i^-(\omega) + 1$ and $Y_i^+(\omega) = Y_i^-(\omega) + 1$. And $X_{i+1}^-(\omega) \geq Y_{i+1}^-(\omega)$ since $X_{i+1}^-(\omega) = X_i^-(\omega) + 1$ and $Y_{i+1}^-(\omega) \leq Y_i^-(\omega)+1$. Case 1.2: If $A_i(\omega) = 0$ and $C_i(\omega) = 0$, $X_i^+(\omega) = X_i^-(\omega) \geq Y_i^-(\omega) = Y_i^+(\omega)$, and $X_{i+1}^-(\omega) = X_i^-(\omega) \geq Y_i^-(\omega) \geq Y_{i+1}^-(\omega)$. Case 1.3: If $A_i(\omega) = 1$ and $C_i(\omega) = 1$, $X_i^+(\omega) \geq Y_i^+(\omega)$ since $X_i^+(\omega) = X_i^-(\omega)+1$ and $Y_i^+(\omega) = Y_i^-(\omega)+1$. And $X_{i+1}^-(\omega) \geq Y_{i+1}^-(\omega)$ since $X_{i+1}^-(\omega) = X_i^-(\omega)$ and $Y_{i+1}^-(\omega) = Y_i^-(\omega)$. Case 1.4: If $A_i = 0$ and $C_i = 1$, $X_i^+(\omega) = X_i^-(\omega) \geq Y_i^-(\omega) = Y_i^+(\omega)$, and $X_{i+1}^-(\omega) \geq Y_{i+1}^-(\omega)$ since $X_{i+1}^-(\omega) = \max\{0, X_i^+(\omega) - 1\}$ and $Y_{i+1}^-(\omega) = \max\{0, Y_i^+(\omega) - 1\}$.

Case 2: $X_i^-(\omega) = \lfloor \bar{n}_2 \rfloor + 1 \geq Y_i^-(\omega)$. Case 2.1: If $A_i(\omega) = 1$ and $C_i(\omega) = 0$, $X_i^+(\omega) = \lfloor \bar{n}_2 \rfloor + 1 \geq Y_i^+(\omega)$, and $X_{i+1}^-(\omega) = \lfloor \bar{n}_2 \rfloor + 1 \geq Y_{i+1}^-(\omega)$. Case 2.2: If $A_i(\omega) = 0$ and $C_i(\omega) = 0$, $X_i^+(\omega) = \lfloor \bar{n}_2 \rfloor + 1 \geq Y_i^+(\omega)$, and $X_{i+1}^-(\omega) = \lfloor \bar{n}_2 \rfloor + 1 \geq Y_{i+1}^-(\omega)$. Case 2.3: If $A_i(\omega) = 1$ and $C_i(\omega) = 1$, $X_i^+(\omega) = \lfloor \bar{n}_2 \rfloor + 1 \geq Y_i^+(\omega)$, and $X_{i+1}^-(\omega) \geq Y_{i+1}^-(\omega)$ since $X_{i+1}^-(\omega) = \lfloor \bar{n}_2 \rfloor$ and $Y_{i+1}^-(\omega) \leq \lfloor \bar{n}_2 \rfloor$. Case 2.4: If $A_i(\omega) = 0$ and $C_i(\omega) = 1$, $X_i^+(\omega) = \lfloor \bar{n}_2 \rfloor + 1 \geq Y_i^+(\omega)$, and $X_{i+1}^-(\omega) \geq Y_{i+1}^-(\omega)$ since $X_{i+1}^-(\omega) = \lfloor \bar{n}_2 \rfloor$ and $Y_{i+1}^-(\omega) \leq \lfloor \bar{n}_2 \rfloor$.

Thus, $X_i^+(\omega) \geq Y_i^+(\omega)$ and $X_{i+1}^-(\omega) \geq Y_{i+1}^-(\omega)$. By induction, $X_t^-(\omega) \geq Y_t^-(\omega)$ and $X_t^+(\omega) \geq Y_t^+(\omega)$ for any t. It implies that $\theta_F(\omega) \leq \theta_N(\omega)$ according to (B.62) and (B.63). Moreover, it is true for any given sample path based on the above analysis.

Let $t_1 \doteq \min\{t \in \mathbb{N} : g_2^{(t_1)}(\alpha_0) < \beta_F(\lfloor n_2 \rfloor)\}$, where $g_2(\cdot)$ is defined in (B.3). We construct a set of samples paths Ω_2 as below. $\omega \in \Omega_2$ if and only if $A_t(\omega) = 1$ and $C_t(\omega) = 0$ for $t \in \{1, 2, \ldots, \lfloor \bar{n}_2 \rfloor + 1\}$, $A_t(\omega) = 0$ and $C_t(\omega) = 0$ for $t \in \{\lfloor \bar{n}_2 \rfloor + 2, \ldots, \lfloor \bar{n}_2 \rfloor + t_1\}$, and $A_t(\omega) = 0$ and $C_t(\omega) = 1$ for $t \in \{\lfloor \bar{n}_2 \rfloor + t_1 + 1, \ldots, 2\lfloor \bar{n}_2 \rfloor + t_1 + 1\}$. Thus, $Y_t^+(\omega) < \lfloor \bar{n}_2 \rfloor + 1 = X_t^+(\omega)$ when $t = \lfloor \bar{n}_2 \rfloor + t_1$, $Y_t^+(\omega) = 1$ and $X_t^+(\omega) = 0$ when $t = 2\lfloor \bar{n}_2 \rfloor + t_1 + 1$.

Based on the above analysis, $I_{\{Y_t^+(\omega)>0\}} \leq I_{\{X_t^+(\omega)>0\}}$ for any t since $Y_t^+(\omega) \leq X_t^+(\omega)$. In addition, $C_t(\omega) I_{\{X_t^+(\omega)>0\}} = 1$ and $C_t(\omega) I_{\{Y_t^+(\omega)>0\}} = 0$ when $t = 2\lfloor \bar{n}_2 \rfloor + t_1 + 1$. Thus, when $\tau \geq 2\lfloor \bar{n}_2 \rfloor + t_1 + 1$, for any $\omega \in \Omega_2$, we have

$$\theta_F(\tau, \omega) = \sum_{t=1}^{\tau} C_t(\omega) I_{\{Y_t^+(\omega)>0\}} / \tau < \sum_{t=1}^{\tau} C_t(\omega) I_{\{X_t^+(\omega)>0\}} / \tau = \theta_N(\tau, \omega).$$

In addition, $\mathbb{P}(\omega \in \Omega_2) > 0$.

Thus, $\theta_F(\tau) = \mathbb{E}_\omega \theta_F(\tau, \omega) < \mathbb{E}_\omega \theta_N(\tau, \omega) = \theta_N(\tau)$ when $\tau \geq 2\lfloor \bar{n}_2 \rfloor + t_1 + 1$. $\square$

Proof of Part (b) under condition (i) in (3.20): When $\alpha_0 = 0$, $V_F(n, 0) = \max\{0, R - \frac{cn}{p_L}\}$ by Lemma 24. Thus, $f(n, 0) = R - \frac{cn}{p_L} = s(n, 0)$ and then there is no difference between the naive policy and the forward-looking policy. $\theta_F = \theta_N$ when $\alpha_0 = 0$. We can focus on the case that $\alpha_0 > 0$. By the definition of $\beta_N(n)$ from (B.55), $\beta_N(n) = 0$ for any $n \leq \bar{n}_1$, where $\bar{n}_1 = \frac{Rp_L}{c}$. It implies that $\beta_F(n) = 0$ since $\beta_F(n) \leq \beta_N(n)$ by Proposition 12. For any $n \geq \bar{n}_1 + 1$, $n > \frac{Rp_H}{c}$ since $\frac{Rp_H}{c} - \frac{Rp_L}{c} < 1$, the forward-looking customer and the naive customer will not join by the proof of Proposition 11.$f(n, 1) = R - c\frac{n}{p_H}$ by the proof of Proposition 13. Thus, in both systems customers join if and only if the number of customers ahead satisfies $n \leq \bar{n}_1$. In addition, forward-looking customers never abandon the service since $\beta_F(n) = \beta_N(n) = 0$ for any $n \leq \bar{n}_1$. It immediately follows that $\theta_F = \theta_N$.

Proof of Part (b) under condition (ii) in (3.20): Similar to part (b) under condition (i), we can focus on the case that $\alpha_0 > 0$. We already know that $\beta_N(n) = 0$ and $\beta_F(n) = 0$ for any $n \leq \bar{n}_1$. For any $n \geq \bar{n}_1 + 1$, $\alpha_0 < \beta_F(\bar{n}_1 + 1) \leq \beta_F(n)$ by Proposition 13-(b) and $\beta_F(n) \leq \beta_N(n)$ by Proposition **??**. Thus, the customers of both types join the system if and only if the number of customers ahead satisfies $n \leq \bar{n}_1$. In addition, forward-looking customers never abandon the service. Thus $\theta_F = \theta_N$. $\square$

B.9 Proof of Proposition 17

Let $\{Z_t^-(\omega), t \in \mathbb{N}\}$ and $\{Z_t^+(\omega), t \in \mathbb{N}\}$ be the number of customers in the system with myopic learners at the beginning of each time period, before arrival and after arrival, respectively, for a given sample path ω. $\{X_t^-(\omega), t \in \mathbb{N}\}$ and $\{X_t^+(\omega), t \in \mathbb{N}\}$ denote the number of naive customers as defined earlier. We know that naive customers and myopic learners have the same belief threshold $\{\beta_M(n), n \in \mathbb{N}\}$ for joining decision.

Next, we will show that $X_t^-(\omega) \geq Z_t^-(\omega)$ and $X_t^+(\omega) \geq Z_t^+(\omega)$ for any $t \in \{1, 2, \dots\}$.

$X_1^-(\omega) = Z_1^-(\omega) = 0$. $X_1^+(\omega) = Z_1^+(\omega) = A_1(\omega)$.

Suppose $X_t^-(\omega) \geq Z_t^-(\omega)$ for $t = 1, 2, \dots, i$. We consider two cases as below.

Case 1: $X_i^-(\omega) = Z_i^-(\omega)$. Then $X_i^+(\omega) = Z_i^+(\omega)$ and $X_{i+1}^-(\omega) \geq Z_{i+1}^-(\omega)$ since myopic learners may abandon the service at the end of time period i. Case 2: $X_i^-(\omega) \geq Z_i^-(\omega) + 1$. Then $X_i^+(\omega) \geq Z_i^+(\omega)$ and $X_{i+1}^-(\omega) \geq Z_{i+1}^-(\omega)$.

Thus, $X_i^+(\omega) \geq Z_i^+(\omega)$ and $X_{i+1}^-(\omega) \geq Z_{i+1}^-(\omega)$. By induction, $X_t^-(\omega) \geq Z_t^-(\omega)$ and $X_t^+(\omega) \geq Z_t^+(\omega)$ for any $t \in \{1, 2, \dots\}$.

For myopic learners, denote the long-run average throughput with given sample path ω by $\theta_M(\omega)$,

$$\theta_M(\omega) = \lim_{k\to\infty} \sum_{t=1}^{k} C_t(\omega) I_{\{Z_t^+(\omega)>0\}}/k \tag{B.65}$$

Recall the average throughput of naive customers from (B.62),

$$\theta_N(\omega) = \lim_{k\to\infty} \sum_{t=1}^{k} C_t(\omega) I_{\{X_t^+(\omega)>0\}}/k \geq \lim_{k\to\infty} \sum_{t=1}^{k} C_t(\omega) I_{\{Z_t^+(\omega)>0\}}/k = \theta_M(\omega).$$

Moreover, it is true for any given sample path based on above analysis. Thus, $\theta_N \geq \theta_M$. □

B.10 Proof of Proposition 18

Proof of Part (a): By the definition of the $V_F(n, y)$ from (3.8),

$$V_F(n, y) \geq V_N(n, y) = R - \left(y\frac{n}{p_H} + (1-y)\frac{n}{p_L} \right) c. \tag{B.66}$$

We also know that

$$V_F(0, y) = R = V_N(0, y). \tag{B.67}$$

According to Proposition 14, for any given y, the threshold for naive policy and forward-looking policy satisfy that $n_S(y) \leq n_F(y)$. Let $\bar{n} = n_S(y)$. When $n \leq \bar{n}$, both the naive customer and the forward-looking customer join the system. Based on (B.66), (B.67), $V_F(n, y)$ is convex decreasing in n according to Proposition 10-(a)(e), and $V_N(n, y)$ is linear decreasing in n, it follows that $V_F(n, y) - V_N(n, y)$ is non-negative and increasing in n when $n \leq \bar{n}$. When $n > \bar{n}$, $V_N(n, y) = 0$ since the naive customer does not join. Thus, $V_F(n, y) - V_N(n, y) = V_F(n, y)$ is decreasing in n by Proposition 10-(a). □

Proof of Part (b): We denote the expected total benefit as $V_F(n, y, R)$ with R as an argument. We will use induction to prove that $V_F(n, y, R)$ is convex in R for any (n, y).

When $n = 0$, $V_F(n, y, R) = R$. It is convex in R for any given (n, y). Suppose $V_F(n, y, R)$ is convex in R when $n = k$ for any given y, we will prove that $V_F(n, y, R)$ is convex in R when $n = k + 1$ for any given y.

If there exists y such that $V_F(k + 1, y, R)$ is not convex in R. It implies that there exists R_1, R_2 and $t \in (0, 1)$ such that $V_F(k + 1, y, tR_1 + (1 - t)R_2) > tV_F(k + 1, y, R_1) + (1 - t)V_F(k + 1, y, R_2)$.

Define

$$d_4 \doteq V_F(k + 1, y, tR_1 + (1 - t)R_2) - (tV_F(k + 1, y, R_1) + (1 - t)V_F(k + 1, y, R_2)). \tag{B.68}$$

Then $d_4 > 0$. Recall the optimality equation for forward-looking customer from (3.9),

$$\begin{aligned}
&V_F(k + 1, y, tR_1 + (1 - t)R_2) \\
&= \max \left\{ 0, -c + (yp_H + (1 - y)p_L)V_F \left(k, \frac{p_H y}{p_H y + p_L(1 - y)}, tR_1 + (1 - t)R_2 \right) \right. \\
&\left. + (y(1 - p_H) + (1 - y)(1 - p_L))V_F \left(k + 1, \frac{(1 - p_H)y}{(1 - p_H)y + (1 - p_L)(1 - y)}, tR_1 + (1 - t)R_2 \right) \right\}.
\end{aligned}$$

Since $V_F(k + 1, y, tR_1 + (1 - t)R_2) \geq d_4 > 0$, then we have

$$\begin{aligned}
&V_F(k + 1, y, tR_1 + (1 - t)R_2) \\
&= - c + (yp_H + (1 - y)p_L)V_F \left(k, \frac{p_H y}{p_H y + p_L(1 - y)}, tR_1 + (1 - t)R_2 \right) \\
&+ (y(1 - p_H) + (1 - y)(1 - p_L))V_F \left(k + 1, \frac{(1 - p_H)y}{(1 - p_H)y + (1 - p_L)(1 - y)}, tR_1 + (1 - t)R_2 \right).
\end{aligned} \tag{B.69}$$

By optimality equation for forward-looking customer, we also have

$$V_F(k+1,y,R_1) \geq -c + (yp_H + (1-y)p_L)V_F\left(k, \frac{p_H y}{p_H y + p_L(1-y)}, R_1\right) + (y(1-p_H) + (1-y)(1-p_L))V_F\left(k+1, \frac{(1-p_H)y}{(1-p_H)y + (1-p_L)(1-y)}, R_1\right), \tag{B.70}$$

$$V_F(k+1,y,R_2) \geq -c + (yp_H + (1-y)p_L)V_F\left(k, \frac{p_H y}{p_H y + p_L(1-y)}, R_2\right) + (y(1-p_H) + (1-y)(1-p_L))V_F\left(k+1, \frac{(1-p_H)y}{(1-p_H)y + (1-p_L)(1-y)}R_2\right). \tag{B.71}$$

By (B.68) through (B.71), it follows that

$$\begin{aligned}
&(yp_H + (1-y)p_L)\left(V_F\left(k, \frac{p_H y}{p_H y + p_L(1-y)}, tR_1 + (1-t)R_2\right) - tV_F\left(k, \frac{p_H y}{p_H y + p_L(1-y)}, R_1\right)\right. \\
&\left. - (1-t)V_F\left(k, \frac{p_H y}{p_H y + p_L(1-y)}, R_2\right)\right) \\
&+(y(1-p_H) + (1-y)(1-p_L))\left(V_F\left(k+1, \frac{(1-p_H)y}{(1-p_H)y + (1-p_L)(1-y)}, tR_1 + (1-t)R_2\right)\right. \\
&-tV_F\left(k+1, \frac{(1-p_H)y}{(1-p_H)y + (1-p_L)(1-y)}R_1\right) \\
&\left. - (1-t)V_F\left(k+1, \frac{(1-p_H)y}{(1-p_H)y + (1-p_L)(1-y)}, R_2\right)\right) \\
&\geq d_4.
\end{aligned}$$

By induction hypothesis,

$$\begin{aligned}
&V_F\left(k, \frac{p_H y}{p_H y + p_L(1-y)}, tR_1 + (1-t)R_2\right) \\
&\leq tV_F\left(k, \frac{p_H y}{p_H y + p_L(1-y)}, R_1\right) + (1-t)V_F\left(k, \frac{p_H y}{p_H y + p_L(1-y)}, R_2\right).
\end{aligned} \tag{B.72}$$

Thus,

$$\begin{aligned}
&V_F\left(k+1, \frac{(1-p_H)y}{(1-p_H)y+(1-p_L)(1-y)}, tR_1+(1-t)R_2\right)\\
-&tV_F\left(k+1, \frac{(1-p_H)y}{(1-p_H)y+(1-p_L)(1-y)}, R_1\right)\\
-&(1-t)V_F\left(k+1, \frac{(1-p_H)y}{(1-p_H)\alpha_0+(1-p_L)(1-y)}, R_2\right)\\
\geq&\frac{d_4}{y(1-p_H)+(1-y)(1-p_L)} \geq \frac{d_4}{1-p_L}.
\end{aligned}$$

Replace y with $\frac{(1-p_H)y}{(1-p_H)y+(1-p_L)(1-y)}$, then

$$V_F(k+1, y, tR_1+(1-t)R_2) - (tV_F(k+1, y, R_1)+(1-t)V_F(k+1, y, R_2)) \geq \frac{d_4}{1-p_L}.$$

Define

$$h_4 \doteq \left\lceil \frac{\ln(d_4/\max\{R_1, R_2\})}{\ln(1-p_L)} \right\rceil + 1. \tag{B.73}$$

Repeat this procedure h_4 times, we can find y such that

$$\begin{aligned}
&V_F(k+1, y, tR_1+(1-t)R_2) - (tV_F(k+1, y, R_1)+(1-t)V_F(k+1, y, R_2))\\
\geq&\frac{d_4}{(1-p_L)^{h_4}} > \max\{R_1, R_2\}.
\end{aligned}$$

Since $V_F(k+1, y, R_1) \geq 0$ and $V_F(k+1, y, R_2) \geq 0$, it immediately follows that

$$V_F(k+1, y, tR_1+(1-t)R_2) > \max\{R_1, R_2\} > tR_1+(1-t)R_2.$$

It contradicts with the fact that $V_F(k+1, y, tR_1+(1-t)R_2) \leq tR_1+(1-t)R_2$. Thus, $V_F(k+1, y, R)$ is convex in R for any y.

By induction, $V_F(n, y, R)$ is convex in R for any (n, y). In addition, we know that $V_F(n, y, R)$ is increasing in R by Proposition 10-(c).

For the naive policy, $V_N(n, y, R) = \max\left\{0, R - \left(\frac{n}{p_H}y + \frac{n}{p_L}(1-y)\right)c\right\}$. It is a piecewise linear and increasing in R. We also know that $V_F(n, y, R) \geq V_N(n, y, R)$ by the definition of the forward-looking

policy. In addition, $V_F(n,y,R) = V_N(n,y,R) = 0$ when $R < \frac{n}{p_H}c$ since both types of customers do not join by Proposition 11 and Proposition 12. And when $R \geq \frac{n}{p_L}c$, $\beta_N(n) = 0$ by (B.56) and $\beta_F(n) = 0$ since $\beta_F(n) \leq \beta_N(n)$ by Proposition 12. It implies that both types of customers will join and the forward-looking customer will never abandon the service when $R \geq \frac{n}{p_L}c$. Thus, $V_F(n,y,R) = V_N(n,y,R)$ when $R \geq \frac{n}{p_L}c$.

Combining the above results, we know that the expected value of learning $V_F(n,y,R) - V_N(n,y,R)$ is increasing and then decreasing in R. □

Proof of Part (c): Similarly, by using c as an argument instead of R, we prove that $V_F(n,y,c)$ is convex in c for any (n,y). We also know that $V_F(n,y,c)$ is decreasing in c by Proposition 10-(d). In addition, $V_N(n,y,c)$ is decreasing and piecewise linear in c, and $V_F(n,y),c) \geq V_N(n,y,c)$ by the definition of the forward-looking policy.

When $c \leq \frac{Rp_L}{n}$, $\beta_N(n) = 0$ by (B.56) and thus $\beta_F(n) = 0$ by Proposition 12. Thus $V_F(n,y,c) = V_N(n,y,c)$ when $c \leq \frac{Rp_L}{n}$ since both types of customers join the system and the forward-looking customer will never abandon. When $c \geq \frac{Rp_H}{n}$, $V_F(n,y,c) = V_N(n,y,c) = 0$ since both types of customers will not join by Proposition 13 and 12.

Based on the above analysis, the expected value of learning $V_F(n,y,c) - V_N(n,y,c)$ is increasing and then decreasing in c. □

B.11 Proof of Proposition 19:

Proof of Part (a): First, we prove two lemmas, which will be used in the remainder of the proof.

Lemma 26. *When $p = p_H$, $U_F^H(n,y) \geq V_F(n,y)$.*

Proof of Lemma 26: Recall the optimality equations for $V_F(n,y)$ and $U_F^H(n,y)$ from (3.9) and (3.22). When $f(n,y) < 0$,

$$V_F(n,y) = 0 \quad \text{and} \quad U_F^H(n,y) = 0. \tag{B.74}$$

When $f(n,y) \geq 0$,

$$V_F(n,y) = -c + (yp_H + (1-y)p_L)V_F(n-1, g_1(y)) + (y(1-p_H) + (1-y)(1-p_L))V_F(n, g_2(y)), \tag{B.75}$$

$$U_F^H(n,y) = -c + p_H U_F^H(n-1, g_1(y)) + (1-p_H)U_F^H(n, g_2(y)), \tag{B.76}$$

where $g_1(y)$ and $g_2(y)$ are defined in (B.2) and (B.3).

We will use induction to prove the claim.

When $n = 0$, $V_F(n,y) = U_F^H(n,y) = R$. $U_F^H(n,y) \geq V_F(n,y)$ holds. Suppose $U_F^H(n,y) \geq V_F(n,y)$ holds for $n = 0, 1, \ldots, k-1$. When $n = k$, suppose there exists y such that $U_F^H(n,y) < V_F(n,y)$.

Define

$$d_5 \doteq V_F(k,y) - U_F^H(k,y). \tag{B.77}$$

Then $d_5 > 0$ and it implies that $f(k,y) \geq 0$. By (B.75) and (B.76), we have

$$\begin{aligned} &-c + (yp_H + (1-y)p_L)V_F(k-1, g_1(y)) + (y(1-p_H) + (1-y)(1-p_L))V_F(k, g_2(y)) - \\ &(-c + p_H U_F^H(k-1, g_1(y)) + (1-p_H)U_F^H(k, g_2(y))) = d_5. \end{aligned} \tag{B.78}$$

Since we know that $V_F(k-1, g_1(y)) \geq V_F(k, g_1(y)) \geq V_F(k, g_2(y))$ since $g_1(y) \geq g_2(y)$ by Proposition 10 and $V_F(k-1, g_1(y)) \leq U_F^H(k-1, g_1(y))$ by induction hypothesis, then we have

$$\begin{aligned} &-c + (yp_H + (1-y)p_L)V_F(k-1, g_1(y)) + (y(1-p_H) + (1-y)(1-p_L))V_F(k, g_2(y)) \\ \leq &-c + p_H V_F(k-1, g_1(y)) + (1-p_H)V_F(k, g_2(y)) \qquad \text{(B.79)} \\ \leq &-c + p_H U_F^H(k-1, g_1(y)) + (1-p_H)V_F(k, g_2(y)). \qquad \text{(B.80)} \end{aligned}$$

Combining (B.78) and (B.80), we have

$$V_F(k, g_2(y)) - U_F^H(k, g_2(y)) \geq \frac{d_5}{1-p_H}.$$

Replace y with $g_2(y)$, then $y \in [0,1]$ and $V_F(k,y) - U_F^H(k,y) \geq \frac{d_5}{1-p_H}$.

Define

$$h_5 \doteq \left\lceil \frac{\ln(d_5/(R+ck/p_H)}{\ln(1-p_H)} \right\rceil + 1. \tag{B.81}$$

Keep repeating the above procedure h_5 times, we get y such that $y \in [0,1]$ and

$$V_F(k,y) - U_F^H(k,y) \geq \frac{d_5}{(1-p_H)^{h_5}} > R + \frac{ck}{p_H}, \tag{B.82}$$

which implies $V_F(k,y) \geq U_F^H(k,y) + \frac{d_5}{(1-p_H)^{h_5}} > -c\frac{k}{p_H} + \frac{d_5}{(1-p_H)^{h_5}} > R$. The reason that $U_F^H(k,y) > -c\frac{k}{p_H}$ is $E(\tau) \leq \frac{k}{p_H}$, where τ is the stopping time of forward-looking customer. It contradicts with the fact that $V_F(k, g_2(y)) \leq R$.

Thus, $U_F^H(k,y) \geq V_F(k,y)$ for any y.

By induction, $U_F^H(n,y) \geq V_F(n,y)$ for any n and $y \in [0,1]$. $\square$

Lemma 27. *Given $p = p_H$, the expected total benefit for the forward-looking customer $U_F^H(n,y)$ is increasing in y.*

Proof of Lemma 27: We will use induction to prove the claim.

When $n = 0$, $U_F^H(n,y) = R$, which is increasing.

Suppose $U_F^H(n,y)$ is increasing in y when $n = 0, 1, \ldots, k-1$. When $n = k$, suppose that there exists y_1 and y_2 such that $y_1 > y_2$ and $U_F^H(n,y_1) < U_F^H(n,y_2)$.

Define

$$d_6 \doteq U_F^H(k,y_2) - U_F^H(k,y_1). \tag{B.83}$$

Then $d_6 > 0$ and it implies that $U_F^H(k,y_2) > 0$ since $U_F^H(k,y_1) \geq V_F(k,y_1) \geq 0$ by Lemma 26. It immediately follows that $f(k,y_2) \geq 0$ by $U_F^H(k,y_2) > 0$ and (3.22). Thus, $f(k,y_1) \geq f(k,y_2) \geq 0$ since $f(k,y)$ is increasing in y by Lemma 25.

Recall the optimality equation given $p = p_H$ from (3.22), we have

$$U_F^H(k, y_1) = -c + p_H U_F^H(k-1, g_1(y_1)) + (1-p_H) U_F^H(k, g_2(y_1)) \tag{B.84}$$

$$\geq -c + p_H U_F^H(k-1, g_1(y_2)) + (1-p_H) U_F^H(k, g_2(y_1)). \tag{B.85}$$

$$U_F^H(k, y_2) = -c + p_H U_F^H(k-1, g_1(y_2)) + (1-p_H) U_F^H(k, g_2(y_2)) \tag{B.86}$$

The inequality (B.84) is because $V_F(k, y_1) > 0$. The inequality (B.85) follows from $g_1(y_1) > g_1(y_2)$ since $g_1(\cdot)$ is an increasing function and $U_F^H(k-1, y)$ is increasing in y by induction hypothesis.

By (B.83), (B.85) and (B.86), we have

$$U_F^H(k, g_2(y_2)) - U_F^H(k, g_2(y_1)) \geq \frac{d_6}{1-p_H}. \tag{B.87}$$

Replace y_1 with $g_2(y_1)$ and y_2 with $g_2(y_2)$, then $y_1, y_2 \in [0,1]$, $y_1 > y_2$ and $U_F^H(k, y_2) - U_F^H(k, y_1) \geq \frac{d_6}{1-p_H}$.

Define

$$h_6 \doteq \left\lceil \frac{\ln(d_6/R)}{\ln(1-p_H)} \right\rceil + 1. \tag{B.88}$$

Keep repeating this procedure h_6 times, we have $y_1, y_2 \in [0,1]$, $y_1 > y_2$ and

$$U_F^H(k, y_2) - U_F^H(k, y_1) \geq \frac{d_6}{(1-p_H)^{h_6}} > R. \tag{B.89}$$

It implies that $U_F^H(k, g_2(y_2)) > R$ since $U_F^H(k, y_1) \geq V_F(k, y_1) \geq 0$ by Lemma 26, which contradicts with the fact that $U_F^H(k, g_2(y_2)) \leq R$. Thus, $U_F^H(k, y_1) \geq U_F^H(k, y_2)$ for any $y_1 > y_2$. $U_F^H(k, y)$ is increasing in y.

By induction, $U_F^H(n, y)$ is increasing in y for any n. $\square$

When $y > \bar{y}_1(n) = \frac{1/p_L - R/cn}{1/p_L - 1/p_H}$, $R - \left(\frac{n}{p_H} y + \frac{n}{p_L}(1-y)\right) c > 0$. Thus, the customer using the naive policy or the myopic policy joins the queue. By Proposition 12, $\beta_N(n) \geq \beta_F(n)$ and thus the forward-looking customer also joins the queue.

Given $p = p_H$, the expected total benefit for the naive customer is $U_N^H(n, y) = R - c\frac{n}{p_H}$ by (3.21).

Since the forward-looking customer also joins, it implies that $f(n,y) \geq 0$ and thus $f(n,1) \geq 0$ because $f(n,y)$ is increasing in y by Lemma 25. Also, $f(k,1) \geq 0$ for $k \in \{0,1,\ldots,n\}$ due to the monotonicity of $f(n,y)$ in n by the proof of Proposition 13-(b). Recall the optimality equation from (3.22), for $k \in \{1,2,\ldots,n\}$, we have

$$U_F^H(k,1) = -c + p_H U_F^H(k-1,1) + (1-p_H)U_F^H(k,1). \tag{B.90}$$

It implies that $U_F^H(k,1) = -\frac{c}{p_H} + U_F^H(k-1,1)$. Since $U_F^H(0,1) = R$, we have

$$U_F^H(n,1) = R - c\frac{n}{p_H}. \tag{B.91}$$

Thus, $U_N^H(n,y) = U_F^H(n,1) \geq U_F^H(n,y)$ since $U_F^H(n,y)$ is increasing in y by Lemma 27.

The loss from learning $U_N^H(n,y) - U_F^H(n,y)$ is decreasing with y since $U_N^H(n,y) = U_F^H(n,1)$, which does not depend on y, and $U_F^H(n,y)$ is increasing with y from Lemma 27.

Next we compare the expected total benefit between the forward-looking policy and the myopic policy. Define

$$T_1 = \min\{t \in \mathbb{N}_+ : \text{the myopic learner gets service at the end of time period t}\}, \tag{B.92}$$

$$T_2 = \min\{t \in \mathbb{N}_+ : \text{the myopic learner abandons the service at the end of time period t}\}. \tag{B.93}$$

When $p = p_H$, we have

$$\begin{aligned} U_F^H(n,y) - U_M^H(n,y) =& \mathbb{E}[U_F^H(n_{T_2}, y_{T_2})|T_2 < T_1]P(T_2 < T_1) \\ \geq& \mathbb{E}[V_F(n_{T_2}, y_{T_2})|T_2 < T_1]P(T_2 < T_1) \qquad \text{(B.94)} \\ \geq& 0 \end{aligned}$$

The inequality (B.94) follows from the fact that $U_F^H(n,y) \geq V_F(n,y)$ for any n and y when $p = p_H$ by Lemma 26.

Thus, $U_F^H(n,y) - U_M^H(n,y) \geq 0$ when $p = p_H$. The forward-looking policy results in larger expected total benefit than the myopic policy given $p = p_H$.

Based on the above analysis, the naive policy results in the maximum expected total benefit when $y > \bar{y}_1(n)$, where $\bar{y}_1(n)$ is defined in (3.26). $\square$

Proof of Part (b): First, we introduce a lemma, which will be used in the remainder of the proof.

Lemma 28. *Given $p = p_L$, the expected total benefit for the forward-looking customer $U_F^L(n, y)$ is decreasing in y.*

Proof of Lemma 28: We will use induction to prove this claim.

When $n = 0$, $U_F^L(n, y) = R$. It is decreasing when $n = 0$.

Suppose $U_F^L(n, y)$ is decreasing in y when $n = 0, 1, \ldots, k-1$. When $n = k$, suppose that there exist $y_1, y_2 \in [0, 1]$ such that $y_1 > y_2$ and $U_F^L(n, y_1) > U_F^L(n, y_2)$.

Define

$$d_7 \doteq U_F^L(k, y_1) - U_F^L(k, y_2). \tag{B.95}$$

Case 1: If $f(k, y_2) \geq 0$, then $f(k, y_1) \geq f(k, y_2) \geq 0$ by Lemma 25 since $y_1 > y_2$. Recall the optimality equation for $U_F^L(n, y)$ from (3.22), we have

$$\begin{aligned} U_F^L(k, y_1) =& -c + p_L U_F^L(k-1, g_1(y_1)) + (1-p_L) U_F^L(k, g_2(y_1)) \\ \leq& -c + p_L U_F^L(k-1, g_1(y_2)) + (1-p_L) U_F^L(k, g_2(y_1)), \end{aligned} \tag{B.96}$$

$$U_F^L(k, y_2) = -c + p_L U_F^L(k-1, g_1(y_2)) + (1-p_L) U_F^L(k, g_2(y_2)). \tag{B.97}$$

The inequality (B.96) follows from the fact that $g_1(y_1) > g_1(y_2)$ since $g_1(y)$ is strictly increasing in y, and $U_F^L(k-1, y)$ is decreasing in y by induction hypothesis.

Combing (B.96) and (B.97), we have

$$U_F^L(k, g_2(y_1)) - U_F^L(k, g_2(y_2)) \geq \frac{d_7}{1-p_L}. \tag{B.98}$$

Case 2: If $f(k, y_2) < 0$, then $U_F^L(k, y_2) = 0$ by (3.22). Thus $U_F^L(k, y_1) = d_7 > 0$ and it implies that $f(k, y_1) \geq 0$.

$$\begin{aligned} U_F^L(k, y_1) =& -c + p_L U_F^L(k-1, g_1(y_1)) + (1-p_L) U_F^L(k, g_2(y_1)) \\ \leq& -c + p_L U_F^L(k-1, g_1(y_2)) + (1-p_L) U_F^L(k, g_2(y_1)) \quad \text{(B.99)} \\ \leq& -c + p_L V_F(k-1, g_1(y_2)) + (1-p_L) U_F^L(k, g_2(y_1)). \quad \text{(B.100)} \end{aligned}$$

The inequality (B.99) follows from $U_F^L(k-1, g_1(y_1)) \leq U_F^L(k-1, g_1(y_2))$ since $U_F^L(k-1, y)$ in decreasing in y by induction and $g_1(y_1) > g_1(y_2)$. The inequality (B.100) is because $U_F^L(k-1, g_1(y_2)) \leq V_F(k-1, g_1(y_2))$ by Lemma 30.

Thus, we have

$$-c + p_L V_F(k-1, g_1(y_2)) + (1-p_L) U_F^L(k, g_2(y_1)) \geq d_7. \quad \text{(B.101)}$$

Since $f(k, y_2) < 0$, we have

$$\begin{aligned} & -c + (y_2 p_H + (1-y_2) p_L) V_F(k-1, g_1(y_2)) + (y_2(1-p_H) + (1-y_2)(1-p_L)) V_F(k, g_2(y_2)) < 0 \\ \Rightarrow & -c + (y_2 p_H + (1-y_2) p_L) V_F(k-1, g_1(y_2)) < 0 \quad \text{(B.102)} \\ \Rightarrow & -c + p_L V_F(k-1, g_1(y_2)) < 0. \quad \text{(B.103)} \end{aligned}$$

Combining (B.101) and (B.103), we have

$$U_F^L(k, g_2(y_1)) \geq \frac{d_7}{1-p_L}. \quad \text{(B.104)}$$

In addition, $f(k, g_2(y_2)) \leq f(k, y_2) = 0$ since $f(k, y)$ is increasing in y by Lemma 25 and $g_2(y_2) \leq y_2$. Thus, $U_F^L(k, g_2(y_2)) = 0$ and we have

$$U_F^L(k, g_2(y_1)) - U_F^L(k, g_2(y_2)) \geq \frac{d_7}{1-p_L}. \quad \text{(B.105)}$$

Thus, in both cases,

$$U_F^L(k, g_2(y_1)) - U_F^L(k, g_2(y_2)) \geq \frac{d_7}{1 - p_L}. \tag{B.106}$$

Replace y_1 with $g_2(y_1)$ and y_2 with $g_2(y_2)$, then $y_1, y_2 \in [0, 1]$, $y_1 > y_2$ and $U_F^L(k, y_2) - U_F^L(k, y_1) \geq \frac{d_7}{1-p_L}$.

Define

$$h_7 \doteq \left\lceil \frac{\ln(d_7/(R + ck/p_L))}{\ln(1 - p_L)} \right\rceil + 1. \tag{B.107}$$

Repeat the above procedure h_7 times, we have $y_1, y_2 \in [0, 1]$, $y_1 > y_2$ and $U_F^L(k, y_1) - U_F^L(k, y_2) \geq \frac{d_7}{(1-p_L)^{h_7}} > R + ck/p_L$.

Then $U_F^L(k, y_1) \geq U_F^L(k, y_2) + \frac{d_7}{(1-p_L)^h} \geq -c\frac{k}{pL} + \frac{d_7}{(1-p_L)^{h_7}} > R$. The reason that $U_F^L(k, y) \geq -c\frac{k}{p_L}$ is $E(\tau) \leq \frac{k}{p_L}$, where τ is the stopping time of the forward-looking customer. It contradicts with the fact that $U_F^L(k, y_1) \leq R$.

Thus, $U_F^L(k, y_1) \leq U_F^L(k, y_2)$ for any $y_1 > y_2$. $U_F^L(k, y)$ is decreasing in y.

By induction, $U_F^L(n, y)$ is decreasing in y for any n. □

When $y > \bar{y}_1(n)$, all types of customers join, and the expected total benefit of naive customer is $U_N^L(n, y) = R - c\frac{n}{p_L}$.

Since the forward-looking customer also joins, it implies that $f(n, y) \geq 0$ and thus $f(n, 1) \geq f(n, y) \geq 0$ by Lemma 25. Also, $f(k, 1) > 0$ for $k = 0, 1, 2, \ldots, n-1$ due to the monotonicity of $f(n, y)$ in n by the proof of Proposition 13-(b). Recall the optimality equation for $U_F^L(n, y)$ from (3.22), for $k \in \{1, 2, \ldots, n\}$, we have

$$U_F^L(k, 1) = -c + p_L U_F^L(k-1, 1) + (1 - p_L) U_F^L(k, 1). \tag{B.108}$$

Then $U_F^L(k, 1) = -\frac{c}{p_L} + U_F^L(k-1, 1)$ when $k = 1, 2, \ldots, n-1$. Since $U_F^L(0, 1) = R$, it follows that

$$U_F^L(n, 1) = R - c\frac{n}{p_L} = U_N^L(n, y). \tag{B.109}$$

Thus, $U_N^L(n, y) \leq U_F^L(n, y)$ since $U_F^L(n, 1) \leq U_F^L(n, y)$ by Lemma 28.

The gain from learning $U_F^L(n,y) - U_F^L(n,1)$ is decreasing with y by Lemma 28.

Next, we compare the forward-looking policy and the myopic policy.

When $p = p_L$, recall the definition of T_1 and T_2 from (B.92) and (B.93), we have:

$$\begin{aligned} U_F^L(n,y) - U_M^L(n,y) =& E[U_F^L(n_{T_2}, y_{T_2})|T_2 < T_1]P(T_2 < T_1) \\ \leq& E[U_F^L(n_{T_2}, 0)|T_2 < T_1]P(T_2 < T_1) \quad \text{(B.110)} \\ \leq& 0 \quad \text{(B.111)} \end{aligned}$$

The inequality (B.110) follows from the fact that $U_F^L(n,y)$ is decreasing in y for any n given $p = p_L$ by Lemma 28. The inequality (B.111) is because $U_F^L(n_{T_2}, 0) \leq 0$ given the fact that the myopic learner abandons the queue. The reason is as following. Recall the optimality equation for forward-looking policy given $p = p_L$ from (3.22). If $f(n_{T_2}, 0) < 0$, then $U_F^L(n_{T_2}, 0) = 0$. If $f(n_{T_2}, 0) \geq 0$, then $f(k, 0) \geq 0$ for any $k \leq n_{T_2}$ by the proof of Proposition 13-(b). Thus, $U_F^L(k,0) = -c + p_L U_F^L(k-1,0) + (1-p_L)U_F^L(k,0)$ for $k \leq n_{T_2}$, which implies that $U_F^L(n_{T_2}, 0) = R - c\frac{n_{T_2}}{p_L}$. Since the myopic learner abandons the service at (n_{T_2}, y_{T_2}), then $R - c\left(y_{T_2}\frac{n_{T_2}}{p_H} + (1-y_{T_2})\frac{n_{T_2}}{p_L}\right) < 0$ by (3.12). Thus, $U_F^L(n_{T_2}, 0) = R - c\frac{n_{T_2}}{p_L} \leq R - c\left(y_{T_2}\frac{n_{T_2}}{p_H} + (1-y_{T_2})\frac{n_{T_2}}{p_L}\right) < 0$.

Thus, $U_F^L(n,y) - U_M^L(n,y) \leq 0$ given $p = p_L$. The forward-looking policy results in smaller expected total benefit than the myopic policy given $p = p_L$.

Based on the above analysis, the myopic policy results in maximum expected total benefit. □

Proof of Part (c): We first prove a lemma, which will be used in the remainder of the proof.

Lemma 29. *When $\beta_N(n) \in (0,1)$, $\beta_F(n) < \beta_N(n)$.*

Proof of Lemma 29: Recall the optimality equation for forward-looking customers from (3.9),

$$\begin{aligned} V_F(n,y) = \max\Bigg\{0, -c + (yp_H + (1-y)p_L)V_F\left(n-1, \frac{p_H y}{p_H y + p_L(1-y)}\right) + \\ (y(1-p_H) + (1-y)(1-p_L))V_F\left(n, \frac{(1-p_H)y}{(1-p_H)y + (1-p_L)(1-y)}\right)\Bigg\}. \quad \text{(B.112)} \end{aligned}$$

Recall the expected total benefit for naive customers from (3.15),

$$V_N(n,y) = \max\left\{0, R - \left(\frac{n}{p_H}y + \frac{n}{p_L}(1-y)\right)c\right\}.$$

Recall $s(n,y)$ from (B.54), $s(n,y) = R - \left(\frac{n}{p_H}y + \frac{n}{p_L}(1-y)\right)c$. By elementary calculations,

$$\begin{aligned} s(n,y) =& -c + (yp_H + (1-y)p_L)s\left(n-1, \frac{p_H y}{p_H y + p_L(1-y)}\right) \\ &+ (y(1-p_H) + (1-y)(1-p_L))s\left(n, \frac{(1-p_H)y}{(1-p_H)y + (1-p_L)(1-y)}\right). \end{aligned}$$

Thus,

$$\begin{aligned} & f(n, \beta_N(n)) \\ &= -c + (\beta_N(n)p_H + (1-\beta_N(n))p_L)V_F\left(n-1, \frac{p_H\beta_N(n)}{p_H\beta_N(n) + p_L(1-\beta_N(n))}\right) \\ &+ (\beta_N(n)(1-p_H) + (1-\beta_N(n))(1-p_L))V_F\left(n, \frac{(1-p_H)\beta_N(n)}{(1-p_H)\beta_N(n) + (1-p_L)(1-\beta_N(n))}\right) \\ &> -c + (\beta_N(n)p_H + (1-\beta_N(n))p_L)s\left(n-1, \frac{p_H\beta_N(n)}{p_H\beta_N(n) + p_L(1-\beta_N(n))}\right) \\ &+ (\beta_N(n)(1-p_H) + (1-\beta_N(n))(1-p_L))s\left(n, \frac{(1-p_H)\beta_N(n)}{(1-p_H)\beta_N(n) + (1-p_L)(1-\beta_N(n))}\right) \quad \text{(B.113)} \\ &= 0. \end{aligned}$$

The inequality (B.113) is because when $\beta_N(n) \in (0,1)$, $\frac{(1-p_H)\beta_N(n)}{(1-p_H)\beta_N(n)+(1-p_L)(1-\beta_N(n))} < \beta_N(n)$ and it implies that

$$\begin{aligned} & s\left(n, \frac{(1-p_H)\beta_N(n)}{(1-p_H)\beta_N(n) + (1-p_L)(1-\beta_N(n))}\right) < s(n, \beta_N(n)) = 0 \\ & \leq V_F\left(n, \frac{(1-p_H)\beta_N(n)}{(1-p_H)\beta_N(n) + (1-p_L)(1-\beta_N(n))}\right), \end{aligned} \quad \text{(B.114)}$$

and by the definition of $V_F(n,y)$ from (3.8),

$$\begin{aligned} V_F\left(n-1, \frac{p_H\beta_N(n)}{p_H\beta_N(n) + p_L(1-\beta_N(n))}\right) &\geq V_N\left(n-1, \frac{p_H\beta_N(n)}{p_H\beta_N(n) + p_L(1-\beta_N(n))}\right) \\ &\geq s\left(n-1, \frac{p_H\beta_N(n)}{p_H\beta_N(n) + p_L(1-\beta_N(n))}\right). \end{aligned}$$

It follows that $\beta_F(n) < \beta_N(n)$ by the definition of $\beta_F(n)$ and $\beta_N(n)$ from (B.52) and (B.55). This completes the proof of Lemma 29. □

By Lemma (29), $\beta_F(n) < \beta_N(n)$ when $\beta_N \in (0,1)$. Recall the definition of $\bar{y}_1$ from (3.26), $\bar{y}_1(n) = \beta_N(n) = \beta_M(n)$. Let $\bar{y}_2(n) = \beta_F(n)$, then the forward-looking customer joins while the customer using naive policy or myopic policy does not join when $\bar{y}_2(n) \leq y < \bar{y}_1(n)$. It implies that $U_N^H(n,y) = 0$, $U_M^H(n,y) = 0$ and $U_F^H(n,y) \geq V_F(n,y) \geq 0$ by Lemma 26. The customer will gain from learning, and the forward-looking policy results in maximum expected total benefit. The gain $U_F^H(n,y)$ is increasing with y by Lemma 27. □

Proof of Part (d): First we prove a lemma, which will be used later.

Lemma 30. *When $p = p_L$, $U_F^L(n,y) \leq V_F(n,y)$.*

Proof of Lemma 30: We will use induction to show the claim.

When $n = 0$, $V_F(n,y) = U_F^L(n,y) = R$. $U_F^L(n,y) \leq V_F(n,y)$ holds. Suppose $U_F^L(n,y) \leq V_F(n,y)$ holds for $n = 0, 1, \ldots, k-1$. When $n = k$, suppose there exists y such that $U_F^L(n,y) > V_F(n,y)$.

Define

$$d_9 \doteq U_F^L(k,y) - V_F(n,y). \tag{B.115}$$

If $f(k,y) < 0$, then $U_F^L(k,y) = 0$ and $V_F(n,y) = 0$ by (3.9) and (3.22). This can not be true.

Thus, $f(k,y) \geq 0$. By the optimality equations from (3.9) and (3.22),

$$\begin{aligned}
V_F(k,y) =& -c + (yp_H + (1-y)p_L)V_F(k-1, g_1(y)) + (y(1-p_H) + (1-y)(1-p_L))V_F(k, g_2(y)) \\
\geq& -c + p_L V_F(k-1, g_1(y)) + (1-p_L)V_F(k, g_2(y)) && \text{(B.116)} \\
\geq& -c + p_L U_F^L(k-1, g_1(y)) + (1-p_L)V_F(k, g_2(y)) && \text{(B.117)} \\
U_F^L(k,y) =& -c + p_L U_F^L(k-1, g_1(y)) + (1-p_L)U_F^L(k, g_2(y)). && \text{(B.118)}
\end{aligned}$$

The inequality (B.116) follows by $V_F(k-1, g_1(y)) \geq V_F(k, g_1(y)) \geq V_F(k, g_2(y))$ since $g_1(y) \geq g_2(y)$ by Proposition 10-(a)(b). The inequality (B.117) is because $U_F^L(k-1, g_1(y)) \leq V_F(k-1, g_1(y))$ by induction hypothesis.

Combining (B.115, (B.117) and (B.118), we have

$$U_F^L(k, g_2(y)) - V_F(k, g_2(y)) \geq \frac{d_9}{1-p_L}. \tag{B.119}$$

Replace y with $g_2(y)$, then $y \in [0,1]$ and

$$U_F^L(k,y) - V_F(k,y) \geq \frac{d_9}{1-p_L}. \tag{B.120}$$

Define

$$h_9 \doteq \left\lceil \frac{\ln(d_9/R)}{\ln(1-p_L)} \right\rceil + 1. \tag{B.121}$$

Keep repeating this procedure h_9 times, we will find y such that $y \in [0,1]$ and

$$U_F^L(k,y) - V_F(k,y) \geq \frac{d_9}{(1-p_L)^{h_9}} > R, \tag{B.122}$$

which implies $U_F^L(k,y) > R$. It contradicts with the fact that $U_F^L(k,y) \leq R$.

Thus, $U_F^L(k,y) \leq V_F(k,y)$ for any y.

By induction, $U_F^L(n,y) \leq V_F(n,y)$ for any n and y. $\square$

When $\bar{y}_2(n) \leq y < \bar{y}_1(n)$, the forward-looking customer joins while the naive customer and the myopic learner do not join. It implies that $f(n,y) \geq 0$ and $R - c\left(y\frac{n}{p_L} + (1-y)\frac{n}{p_L}\right) < 0$.

Next, we prove that $U_F^L(n,0) = V_F(n,0) = 0$.

By Lemma 24,

$$V_F(n,0) = \max\left\{0, R - c\frac{n}{p_L}\right\}. \tag{B.123}$$

Suppose $V_F(n,0) > 0$, then $V_F(n,0) \leq R - c(y\frac{n}{p_H} + (1-y)\frac{n}{p_L} < 0$, which contradicts with that $V_F(n,0) > 0$. Thus, $V_F(n,0) = 0$ and $U_F^L(n,0) \leq V_F(n,0) = 0$ by Lemma 30. Then $U_F^L(n,y) \leq U_F^L(n,0) = 0$ since $U_F^L(n,y)$ is decreasing in y by Lemma 28. Thus, the naive policy and myopic policy result in maximum expected total benefit, and the value of learning is non-positive. The loss from learning is $-U_F^L(n,y)$ is increasing in y by Lemma 28. $\square$

BIBLIOGRAPHY

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

Afeche, P. and Sarhangian, V. (2015). Rational abandonment from priority queues: equilibrium strategy and pricing implications.

Aksin, Z., Armony, M., and Mehrotra, V. (2007). The modern call center: A multi-disciplinary perspective on operations management research. *Production and operations management*, 16(6):665–688.

Andradóttir, S., Ayhan, H., and Down, D. G. (2017). Resource pooling in the presence of failures: Efficiency versus risk. *European Journal of Operational Research*, 256(1):230–241.

Armbrüster, T. (2006). *The Economics and Sociology of Management Consulting*. Cambridge University Press.

Armony, M., Roels, G., and Song, H. (2017). Pooling queues with discretionary service capacity. Working Paper.

Armony, M., Shimkin, N., and Whitt, W. (2009). The impact of delay announcements in many-server queues with abandonment. *Operations Research*, 57(1):66–81.

Ata, B. and Peng, X. (2017). An equilibrium analysis of a multiclass queue with endogenous abandonments in heavy traffic. *Operations Research*, 66(1):163–183.

Aviv, Y. and Pazgal, A. (2008). Optimal pricing of seasonal products in the presence of forward-looking consumers. *Manufacturing & Service Operations Management*, 10(3):339–359.

Azziz, R. (2014). Implementing shared services in higher education. `https://www.universitybusiness.com/article/shared-services`, Accessed on 7/8/2018.

Benjaafar, S. (1995). Performance bounds for the effectiveness of pooling in multi-processing systems. *European Journal of Operational Research*, 87(2):375–388.

Besanko, D. and Winston, W. L. (1990). Optimal price skimming by a monopolist facing rational consumers. *Management science*, 36(5):555–567.

Bondarouk, T. (2014). *Shared Services as a New Organizational Form*. Emerald Group Publishing.

Cachon, G. P. and Swinney, R. (2009). Purchasing, pricing, and quick response in the presence of strategic consumers. *Management Science*, 55(3):497–511.

Calabrese, J. M. (1992). Optimal workload allocation in open networks of multiserver queues. *Management Science*, 38(12):1792–1802.

Cattani, K. and Schmidt, G. M. (2005). The pooling principle. *INFORMS Transactions on Education*, 5(2):17–24.

Coase, R. H. (1972). Durability and monopoly. *The Journal of Law and Economics*, 15(1):143–149.

Cui, S. and Veeraraghavan, S. (2016). Blind queues: The impact of consumer beliefs on revenues and congestion. *Management Science*, 62(12):3656–3672.

Debo, L. and Veeraraghavan, S. (2014). Equilibrium in queues under unknown service times and service value. *Operations Research*, 62(1):38–57.

Do, H., Shunko, M., Lucas, M. T., and Novak, D. (2015). On the pooling of queues: How server behavior affects performance. Working Paper.

Edelson, N. M. and Hilderbrand, D. K. (1975). Congestion Tolls for Poisson Queuing Processes. *Econometrica*, 43(1):81–92.

Emadi, S. M. and Swaminathan, J. M. (2017). Impact of callers history on abandonment: Model and implications. Technical report, working paper.

Flatto, L. and McKean, H. (1977). Two queues in parallel. *Communications on Pure and Applied Mathematics*, 30(2):255–263.

Gai, Y., Liu, H., and Krishnamachari, B. (2011). A packet dropping mechanism for efficient operation of M/M/1 queues with selfish users. In *2011 Proceedings IEEE INFOCOM*, pages 2687–2695.

Gans, N., Koole, G., and Mandelbaum, A. (2003a). Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management*, 5(2):79–141.

Gans, N., Koole, G., and Mandelbaum, A. (2003b). Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management*, 5(2):79–141.

Garnett, O., Mandelbaum, A., and Reiman, M. (2002). Designing a call center with impatient customers. *Manufacturing & Service Operations Management*, 4(3):208–227.

Gilbert, S. M. and Weng, Z. K. (1998). Incentive Effects Favor Nonconsolidating Queues in a Service System: The Principal-Agent Perspective. *Management Science*, 44(12-part-1):1662–1669.

Gilboa-Freedman, G., Hassin, R., and Kerner, Y. (2014). The price of anarchy in the Markovian single server queue. *IEEE Transactions on Automatic Control*, 59(2):455–459.

Grassmann, W. K. (1980). Transient and steady state results for two parallel queues. *Omega*, 8(1):105–112.

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

Gupta, V., Harchol Balter, M., Sigman, K., and Whitt, W. (2007). Analysis of Join-the-shortest-queue Routing for Web Server Farms. *Perform. Eval.*, 64(9-12):1062–1081.

Haight, F. A. (1958). Two queues in parallel. *Biometrika*, 45(3-4):401–410.

Hassin, R. (1985). On the optimality of first come last served queues. *Econometrica*, 53(1):201–02.

Hassin, R. (1986). Consumer information in markets with random product quality: The case of queues and balking. *Econometrica: Journal of the Econometric Society*, 54(5):1185–1195.

Hassin, R. (2016a). Rational Queueing. (Chapman & Hall/CRC Series in Operations Research) 1st Edition.

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

Hassin, R. and Haviv, M. (1995). Equilibrium strategies for queues with impatient customers. *Operations Research Letters*, 17(1):41–45.

Hassin, R. and Haviv, M. (2003). *To Queue or Not To Queue: Equilibrium Behavior in Queueing Systems*, volume 59. Springer Science & Business Media.

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

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

Haviv, M. and Oz, B. (2016). Regulating an observable M/M/1 queue. *Operations Research Letters*, 44(2):196–198.

Hong, L. J., Xu, X., and Zhang, S. H. (2015). Capacity reservation for time-sensitive service providers: An application in seaport management. *European Journal of Operational Research*, 245(2):470–479.

Hu, M., Li, Y., and 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.

Jouini, O., Akşin, Z., and Dallery, Y. (2011). Call centers with delay information: Models and insights. *Manufacturing & Service Operations Management*, 13(4):534–548.

Jouini, O., Dallery, Y., and Akşin, Z. (2009). Queueing models for full-flexible multi-class call centers with real-time anticipated delays. *International Journal of Production Economics*, 120(2):389–399.

Jouini, O., Dallery, Y., and Nait-Abdallah, R. (2008). Analysis of the impact of team-based organizations in call center management. *Management Science*, 54(2):400–414.

Karacostas, C. (2018). Voting wait time at FAC for primaries 14 times longer than Travis County as a whole. `http://www.dailytexanonline.com/2018/03/22/`, Accessed on 7/1/2018.

Kingman, J. F. (1961). Two similar queues in parallel. *The Annals of Mathematical Statistics*, 32(4):1314–1323.

Kulkarni, V. G. (2010). *Modeling and Analysis of Stochastic Systems*. CRC Press.

Kuzu, K. (2015). Comparisons of perceptions and behavior in ticket queues and physical queues. *Service Science*, 7(4):294–314.

Kuzu, K., Gao, L., and Xu, S. H. (2017). To wait or not to wait: The theory and practice of ticket queues.

Li, J., Granados, N., and Netessine, S. (2014). Are consumers strategic? structural estimation from the air-travel industry. *Management Science*, 60(9):2114–2137.

Lin, Y.-T., Parlaktürk, A. K., and Swaminathan, J. M. (2018). Are strategic customers bad for a supply chain? *Manufacturing & Service Operations Management*.

Littlechild, S. (1974). Optimal arrival rate in a simple queueing system. *International Journal of Production Research*, 12(3):391–397.

Liu, Q. and Zhang, D. (2013). Dynamic pricing competition with strategic customers under vertical product differentiation. *Management Science*, 59(1):84–101.

Lu, Y., Musalem, A., Olivares, M., and Schilkrut, A. (2013). Measuring the effect of queues on customer purchases. *Management Science*, 59(8):1743–1763.

Mader, D. and Roth, D. T. (2015). Scaling implementation of shared services. https://obamawhitehouse.archives.gov/blog/2015/10/22/scaling-implementation-shared-services, Accessed on 7/8/2018.

Mandelbaum, A. and Reiman, M. I. (1998). On pooling in queueing networks. *Management Science*, 44(7):971–981.

Mandelbaum, A. and Shimkin, N. (2000). A model for rational abandonments from invisible queues. *Queueing Systems*, 36(1-3):141–173.

Mendelson, H. (1985). Pricing computer services: queueing effects. *Communications of the ACM*, 28(3):312–321.

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

Nelson, R. D. and Philips, T. K. (1993). An approximation for the mean response time for shortest queue routing with general interarrival and service times. *Performance evaluation*, 17(2):123–139.

Parlaktürk, A. K. (2012). The value of product variety when selling to strategic consumers. *Manufacturing & Service Operations Management*, 14(3):371–385.

Pender, J. and Jennings, O. (2015). Comparisons of standard and ticket queues. *arXiv preprint arXiv:1505.00819*.

Rao, B. and Posner, M. (1987). Algorithmic and approximation analyses of the shorter queue model. *Naval Research Logistics (NRL)*, 34(3):381–398.

Ravner, L. and Shamir, N. (2017). Pricing strategy, capacity level and collusion in a market with delay sensitivity. Working Paper.

Rodriguez, S. (2014). Verizon wireless to close five call centers, consolidate seven others. *The Los Angeles Times*. http://articles.latimes.com/2014/feb/12/business/la-fi-tn-verizon-call-centers-moving-employees-20140212,Accessed on 7/1/2018.

Ros, D. and Tuffin, B. (2004). A mathematical model of the Paris metro pricing scheme for charging packet networks. *Computer Networks*, 46(1):73–85.

Rothkopf, M. H. and Rech, P. (1987). Perspectives on queues: Combining queues is not always beneficial. *Operations Research*, 35(6):906–909.

Schmidt, J. (1997). Breaking down fiefdoms. *Management Review*, 86(1):45–49.

Selen, J., Adan, I., Kapodistria, S., and van Leeuwaarden, J. (2016). Steady-state analysis of shortest expected delay routing. *Queueing Systems*, 84(3-4):309–354.

Shunko, M., Niederhoff, J., and Rosokha, Y. (2018). Humans Are Not Machines: The Behavioral Impact of Queueing Design on Service Time. *Management Science*, 64(1):453–473.

Smith, D. R. and Whitt, W. (1981). Resource sharing for efficiency in traffic systems. *Bell System Technical Journal*, 60(1):39–55.

Song, H., Tucker, A. L., and Murrell, K. L. (2015). The diseconomies of queue pooling: An empirical investigation of emergency department length of stay. *Management Science*, 61(12):3032–3053.

Southwest (2012). Investor Relations: Southwest Airlines Announces Reservations Consolidation Into New Atlanta Call Center. http://investors.southwest.com/news-and-events/news-releases/2012/28-11-2012,Accessed on 7/1/2018.

Stokey, N. L. (1979). Intertemporal price discrimination. *The Quarterly Journal of Economics*, pages 355–371.

Su, X. (2007). Intertemporal pricing with strategic customer behavior. *Management Science*, 53(5):726–741.

Su, X. and Zhang, F. (2008). Strategic customer behavior, commitment, and supply chain performance. *Management Science*, 54(10):1759–1773.

Swinney, R. (2011). Selling to strategic consumers when product value is uncertain: The value of matching supply and demand. *Management Science*, 57(10):1737–1751.

Sztrik, J. (2012). Basic queueing theory. *University of Debrecen, Faculty of Informatics*, 193.

Campbell Public Affairs Institute (2017). Considering Shared Government Services in New York State. https://www.maxwell.syr.edu/uploadedFiles/campbell/NYS-Shared-Services-Guide.pdf, Maxwell School of Citizenship and Public Affairs, Syracuse University, Accessed on 7/8/2018.

The Financial Times (2015). Big ships leave ports awash with problems. www.ft.com/content/ce9a61e0-8705-11e4-982e-00144feabdc0, Accessed on 7/1/2017.

The NYTimes (2016). Why Long Voting Lines Could Have Long-Term Consequences. https://www.nytimes.com/2016/11/09/upshot/why-long-voting-lines-today-could-have-long-term-consequences.html, Accessed on 7/1/2017.

U.S. Department of the Treasury (2017). Shared services. https://www.fiscal.treasury.gov/fsservices/gov/fit/fit_fssp.htm, Accessed on 7/8/2018.

UAFS (2018). Marketing & Communication Toolbox - Design Services. https://uafs.edu/marketing/design-services, Accessed on 7/8/2018.

van Dijk, N. and van der Sluis, E. (2009). Pooling is not the answer. *European Journal of Operational Research*, 197(1):415–421.

van Dijk, N. M. and van der Sluis, E. (2008). To pool or not to pool in call centers. *Production and Operations Management*, 17(3):296–305.

Veeraraghavan, S., Xiao, L., and Zhang, H. (2018). Impatience and learning in queues. *Under Review*.

Wang, J. and Zhou, Y.-P. (2017). Impact of queue configuration on service time: Evidence from a supermarket. *Management Science*, 64(7):2973–3468.

Ward, A. R. (2012). Asymptotic analysis of queueing systems with reneging: A survey of results for fifo, single class models. *Surveys in Operations Research and Management Science*, 17(1):1–14.

Whitt, W. (1999). Improving service by informing customers about anticipated delays. *Management science*, 45(2):192–207.

Whitt, W. (2004). Efficiency-driven heavy-traffic approximations for many-server queues with abandonments. *Management Science*, 50(10):1449–1461.

Xerox (2013). Contact Center Consolidation A Best Practices Blueprint. `https://www.xerox.com/downloads/services/brochure/contact-center-consolidation.pdf`, Accessed on 7/1/2018.

Yang, L., Debo, L., and Gupta, V. (2017). Trading Time in a Congested Environment. *Management Science*, 63(7):2377–2395.

Yu, Q., Allon, G., and Bassamboo, A. (2016). How do delay announcements shape customer behavior? an empirical study. *Management Science*, 63(1):1–20.

Yu, Y., Benjaafar, S., and Gerchak, Y. (2015). Capacity Sharing and Cost Allocation among Independent Firms with Congestion. *Production and Operations Management*, 24(8):1285–1310.

Zohar, E., Mandelbaum, A., and Shimkin, N. (2002). Adaptive behavior of impatient customers in tele-queues: Theory and empirical support. *Management Science*, 48(4):566–583.