Competition vs. Cooperation of Public and Private Healthcare Centres

by

Seyedeh Saloumeh Sadeghzadeh

A Thesis Submitted to the
Graduate School of Sciences and Engineering
in Partial Fulfillment of the Requirements for
the Degree of

Master of Science

in

Industrial Engineering

Koç University

July, 2015

Koç University

Graduate School of Sciences and Engineering

This is to certify that I have examined this copy of a M.Sc. thesis by

Seyedeh Saloumeh Sadeghzadeh

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Committee Members:

Assistant Prof. Pelin G. Canbolat (Advisor)

Associate Prof. Lerzan E. Örmeci (Advisor)

Prof. Zeynep Akşin Karaesmen

Assistant Prof. Bora Çekyay

Prof. Fikri Karaesmen

Date: _____

To my family...

ABSTRACT

In this thesis, we consider a healthcare system consisting of two healthcare centres where one represents a public and the other a private service centre. Each centre is modelled as an M/M/1 queue. Arriving patients have to be served at exactly one of the two centres, provided that buffer sizes are infinite. We consider the waiting cost per time unit spent in the system is a constant, and it is the same for both centres. Furthermore, patients who join the private centre have to pay a fee, while the public centre is free. We examine the system for decentralized and centralized settings when the queues are either observable or unobservable. In the decentralized setting, individuals act in order to minimize their own expected cost, while in the centralized setting, there is a central authority which sends individuals to the servers in order to minimize the expected cost of the whole system.

In the observable case in which the individuals know the length of the queues, we prove that the optimal strategy is of a threshold type for both decentralized and centralized settings. We derive mathematical formulas to calculate the observable decentralized thresholds. A relationship between the thresholds in decentralized and centralized settings is also derived.

In the unobservable case in which the queue length is not known, we prove the existence of a unique symmetric Nash equilibrium for the decentralized case, and obtain explicit expressions to find the optimal policies for both decentralized and centralized settings. Finally, the results are extended to the systems which have finite waiting room capacity. We implement our results numerically where MRI is considered as the health service. The numerical part also included a sensitivity analysis on some performance measures with respect to different parameters of our model.

ÖZETÇE

Bu tezde bir devlet hastanesi ve bir özel hastane olmak üzere iki sağlık merkezinden oluşan bir sağlık sistemini dikkate alıyoruz. Her merkez M/M/1 kuyruk modeli ile modellenmektedir. Buffer Kapasitesi sonsuz olmak şartıyla gelen hastalara sadece bir merkez tarafından hizmet verilmelidir. Sistemde birim zaman bekleme maliyetinin sabit ve iki sistem için de aynı olduğunu dikkate alıyoruz. Buna ek olarak, devlet hastanelerinin ücretsiz olmasına karşın, özel hastaneye katılan hastalar bir ücret ödemek zorundadır. Kuyrukların gözlenebilir ve gözlenemez durumları için sistemi merkezi olmayan ve merkezi ortamlarda incelemekteyiz. Merkezi olmayan ortamda, bireyler kendi beklenen maliyetlerini en aza indirmeye yönelik davranırken, merkezi ortamda tüm sistemin beklenen maliyetini en az indirmek için bireyleri hizmet birimine gönderen merkezi bir otorite mevcuttur.

Bireylerin kuyrukların uzunluğunu bildiği gözlenebilir durumda, hem merkezi olmayan ortam hem de merkezi ortam için en iyi stratejinin eşik değere bağlı strateji tipi olduğunu kanıtlıyoruz. Gözlenebilir ve merkezi olmayan eşik değerlerini hesaplamak için matematiksel formüller türetiyoruz. Merkezi olmayan ve merkezi ortamlarda eşik değerleri arasındaki ilişki için de formüller elde ediyoruz.

Kuyruk uzunluğunun bilinmediği gözlemlenemeyen durumda eşsiz ve simetrik bir Nash dengesinin varlığını merkezi olmayan durum için kanıtlamakta ve merkezi olmayan ve merkezi ortamda en iyi hareket tarzını bulmak için açık ifadeler elde etmekteyiz.

Son olarak, sonuçlar sonlu bekleme odası kapasiteli sistemler için genişletilmektedir. Sonuçlarımızı sayısal çözümleme ile sağlık sisteminin MRI olarak göz önünde bulundurulduğu duruma uyguluyoruz. Sayısal çözümleme kısmı, modelimizdeki farklı parametrelere göre performans ölçütlerine duyarlılık analizi de içermektedir.

ACKNOWLEDGMENTS

I would like to express my deep gratitude towards my advisors, Lerzan E. Örmeci and Pelin G. Canbolat for all their aspiring guidance, useful comments and friendly advice during this study. I have been so fortunate to have such advisors who gave me the freedom to explore on my own, and at the same time the guidance to find the right way. My sincere appreciation is extended to my committee members: Fikri Karaesmen, Zeynep Akşin Karaesmen, and Bora Çekyay. Special thanks to Apostolos Burnetas to share his unpublished results with us.

I would like to appreciate my family for all their unconditional love, concern and support throughout my life. I would also like to deeply thank my friends for their encouragement and support throughout entire process. Finally, I am grateful to one and all, who directly or indirectly, made this work possible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

In several countries the healthcare system includes both public and private sectors. Public services are usually less expensive, but incur a higher waiting time on patients. Moreover, private sectors usually provide more quality in comparison to the public ones. So there is a trade-off for patients between cost and the combination of time and quality.

Healthcare systems are varied in practice. Hospitals, clinics, specialists' offices, and diagnosis centres are a few examples of healthcare systems. In this work, we focus on the effect of the waiting time, where we assume that both centres provide the same quality of services. This assumption restricts the application of our results mostly to diagnosis services. We consider one public and one private diagnosis centre each with a single server and with the same quality of service. We measure the displeasure of waiting by assuming a fixed cost for each unit of time patients spend in the system. The public centre is assumed to provide a free service while to obtain service from the private centre, patients have to pay a fee. Buffer sizes are infinite and as balking is not allowed, patients have to be served at exactly one of the two centres.

We analyse the system from two perspectives. First, when patients decide which system to join in order to minimize their own expected cost, which is called "decentralized setting." Second, when a central authority routes patients to public or private systems in order to minimize the total expected cost of the system, which is called "centralized setting."

Many service systems can be considered as queueing type systems. Healthcare systems are good examples of this type. In a first-come, first-served (FCFS) healthcare system,

patients have to set an appointment and wait until the previous patients finish service. We may have different number of servers, different queues with different capacities, and all other factors which can be modelled in queueing systems. We can have both observable and unobservable queues. In observable queues, the number of individuals in the queue is known to everyone, while in unobservable queues, one is not informed about the length of the queue before joining it. In healthcare systems many queues are virtual queues in which individuals wait at home or other places rather than waiting in a physical queue, so having an observable queue may be hard in practice.

In this study, we examine how the performance of a queueing system changes if we set a central control to send individuals to different servers, instead of allowing them to choose individually. A strategic individual in a service system decides in order to maximize her own pay-off. Her action can affect other individuals' strategy. The case that all individuals in the system act their best responses to others' strategies is called a Nash equilibrium. The equilibrium behaviour of individuals versus the control of a central authority in the system is a rich topic in queueing models. There are many papers in this field. One of the most well known papers is the work of Naor [1] who examines an observable M/M/1 queue and compares decentralized and centralized optimal strategies, showing that in the decentralized optimal strategy, individuals have negative externalities on others.

We also analyse the system from another point of view: the level of information. The comparison between the efficiency of observable queues versus unobservable ones is another rich topic in queueing models. To what extent can the efficiency of the system be influenced by the level of information individuals may have? Is it always better to have more information? The book of Hassin and Haviv "To queue or not to queue" [2] is a comprehensive review on different queueing systems related to the equilibrium behaviour of individuals in both observable and unobservable queues.

In the observable case, the optimality of a threshold type policy is proved both for decentralized and centralized settings. Analysing the equilibrium behaviour of individuals, we obtain mathematical formulas to calculate the public and private

thresholds in the observable decentralized case. Then we compare the thresholds of the individually optimal policy (decentralized) and those of socially optimal policy (centralized). In the unobservable case, we prove the existence of a unique symmetric Nash equilibrium in the decentralized case, and obtain explicit expressions to find the optimal policy, both in decentralized and centralized settings.

The results are extended to the model with finite buffer. Finally, we illustrate our results by using an example based on real data, and perform sensitivity analysis on different performance measures with respect to different parameters.

The rest of this thesis is organized as follows. Chapter 2 presents the related literature in this field. Chapter 3 describes the model and our assumptions. Chapter 4 discusses both decentralized and centralized policies for the observable queues, and Chapter 5 does the same for the unobservable queues. Chapter 6 shows our results via a numerical example of two MRI centres. Finally, Chapter 7 presents our conclusions.

Chapter 2

LITERATURE REVIEW

Healthcare systems are among the most complicated systems. In 2012, yearly healthcare related costs were at least $6.5 trillion (according to statistics from the "World Health Organization"), or about 9.6 percent of global gross domestic product (GDP), which makes it one of the largest industries in the global economy. On average, the public share in healthcare is about 60 percent. This spending has a big influence on the health outcomes. Besides spending large amounts of money, healthcare systems are often inefficient and need reforms. Many of the problems in healthcare systems are caused by misguided public intervention [3].

In this thesis, we analyse a simple model of a healthcare system in which a public and a private centre coexist. We analyse the effect of having a central control to govern both centres in comparison to allowing patients choose their servers individually. This study can also help to compare the effects of putting more subsidy on the private centre, in comparison to improving the public one.

The related literature is classified to four categories.

2.1 Mixed Healthcare Systems

The use of public and private health services has been studied in many different literature. There are some papers which consider this problem from an economical point of view. Regidor et al. [4] calculate the percentage of the use of public and private healthcare services, such as general practitioner, specialist and hospital care, according to three socio-economic criteria: educational level, social class, and income. They use data from a sample of 18,837 Spanish subjects. They show that public general practitioners and hospital services tend to favour the lower socio-economic

groups, while there are no socio-economic differences in the specialist visit. The inequities indicate an overuse of public services, or the willingness of people in high socio-economic groups to use private services more.

Basu et al. [5] perform a review of research studies related to the performance of public and private sector delivery in low and middle income countries. Their evaluation does not support the traditional claim that the private sector is usually more efficient than the public sector. Each system has its own strengths and weaknesses.

Buying private health insurance is also another factor which can be influenced by many other factors in a healthcare system. Private insurance can also be seen as a means to reduce the overuse of public healthcare services. Canta and Leroux [6] study a healthcare system in which individuals with different incomes can choose between public and private centres. The public centre is less expensive but it incurs a waiting time depending on the number of people in the centre. In their study, they assume that by purchasing the private insurance, individuals can reach waiting-free service. Their objective is to find the optimal income taxation policy, both when the queue is observable and when it is unobservable. They analyse the case in which the social planner assigns agents to the public system, and also the case in which individuals can choose between public and private options. They show that at the laissez-faire the number of agents joining the public system is much higher in comparison with the first-best optimum, which comes from the externality of individuals' choice on the waiting time. They also analyse the second best allocation when there is a social planner which assigns agents to the public system. According to the first best, it is more probable for low-income agents to be assigned to the public system, compared to their high-income counterparts. In the case that there is no social planner, Canta and Leroux show that only a linear income taxation with a subsidy on private insurance can be used. They prove that in this case, if waiting times are not too high, the optimal policy is that high-income agents pay a tax on the use of private facilities, so they would redistribute resources toward low-income agents. For high waiting times, encouraging individuals to buy private insurance is optimal.

Another paper related to the effect of waiting time on the private insurance demand is the work of Bonet [7]. He shows that the decision of purchasing private insurance, depends on the quality of service in public agents, as well as the customer's income, socio-demographic characteristics and health status. For simplicity, Bonet measures the quality of servers by the length of the time patients have to wait in the system. Results indicate that as the quality of the public healthcare increases (waiting time decreases), the probability of purchasing private health insurance decreases.
Johar et al. [8] also analyse the effect of waiting time on the decision to buy private health insurance. They note that insurance demand is influenced by waiting time, not waiting list. Waiting lists may represent technological advances which allow more procedures to be done in a specific period of time. They extend their work [9] analysing the effect of waiting time on the private insurance demand at the individual level, modelling the individuals' expected waiting cost as a function of their own conditions. The data sets they use for modelling are from the NHS 2004-2005 and the NSW IWT data 2004-2005. They show that it is the high probability of waiting for a long time, not the expected waiting time, which increases the probability of buying the private insurance. They claim that waiting time has no significant impact on the insurance demand.

2.2 Observable Centralized Healthcare Models Analysed by MDP

There are many papers which use Markov decision process to analyse a healthcare system. In these studies, the queues are observable and the objective is to maximize the long run average or the discounted pay-off.
Hajek [10] studies the system of two interactive service stations (station 1 and 2), in which the number of individuals in the system is known by everyone. The arrivals of station 1 and station 2 are Poisson with rates λ_1 and λ_2, respectively. There is another Poisson stream of rate λ. An arriving individual at time t is routed to station 1 with probability $a(t)$ and to station 2 with $1 - a(t)$. Both stations' service time are exponential with rate μ_1 and μ_2. There is also an extra server with exponential

rate μ which spends a proportion of service $d(t)$ to station 1 and proportion $1 - d(t)$ to station 2. The individual departs from the system after being served by one of these three servers. There are two exponential servers with rate ν_1 and ν_2 at station 1 and 2. If these servers complete their service, the individual is routed to the other station with probability $r_{12}(t)$ for station 1, and $r_{21}(t)$ for station 2, and otherwise it remains in the same station. Hajek establishes the existence of optimal controls in a Markovian network with two service stations and linear costs. The optimal controls have a switching curve structure in both finite horizon and long run average cost problems.

The paper also analyses some special cases to obtain previously studied models. When λ_1, λ_2, and μ are the only non-zero parameters, then the model turns into the service priority problem. Harrison [11] shows that the optimal policy of a service priority problem, even with non-exponential service, is a fixed priority policy. There are other similar studies by Gittens [12], Nash [13], Whittle [14], and Varaiya et al. [15]. The related problem in which two server works on a single queue is analysed by Lin and Kumar [16].

If λ_1, ν_1, and μ_1 are the only non-zero parameters, then the model turns into a control problem for tandem queue. Sobel [17] has proved that in such systems the long run average cost is minimized by a full-service policy, even for more general networks, when the service time distribution is non-exponential, and for arbitrary arrival processes, when the holding cost per customer is equal in both queues. Rosberg et al. [13] prove the existence of an optimal policy with switching structure, when the holding cost per customer in the second queue is larger.

Finally, if λ, λ_1, λ_2, μ_1, and μ_2 are the only non-zero parameters, then the model turns into a routing problem. Assuming equal service-time distribution for both stations, Winston [18] and Weber [19] have shown that joining the shortest queue is the optimal policy.

The centralized observable case of our work could be considered as a special case of this model when λ, μ_1, and μ_2 are the only non-zero parameters, if there was no fee

for joining the private centre. Even in our model in which there is a fee for joining the private centre, we prove the existence of an optimal strategy and show that this strategy is of a threshold-type.

2.3 Equilibrium Behaviour of Individuals and Socially Optimal Behaviour

Equilibrium behaviour of customers can substantially influence the efficiency of a queueing system, so it is important to analyse the system based on each individual's attitude.

An alternative viewpoint is to assign a central authority to determine the policy in order to maximize the pay-off for the whole system. The degree of central control is an important factor in many service systems. We can analyse the effect of allowing individuals to choose service providers by calculating the *price of anarchy*, which quantifies the inefficiency created by choice. Price of anarchy is defined as the ratio of the cost of the worst possible Nash equilibrium to the cost of the social optimum solution. All these factors could be analysed both in observable and unobservable queues.

Equilibrium behaviour in different queueing models are studied in Hassin and Haviv's book [2]. The book is a comprehensive survey which contains related literature in this field. It also classifies different queueing models, identifies their results and analyses how they relate to each other.

For the observable case of our study, a related work is Naor's paper [1] in which he analyses a queueing model where customers arrive in a Poisson stream at a service station with exponential service time. The queue is observable, and each customer is aware of the monetary reward of getting service, and also the waiting cost per unit of time. A customer has two alternatives: joining the queue, or balking. Naor analyses both individual and social optimization. In the individual optimization case, each customer decides to join the queue or balk, only based on maximizing her own pay-off. In the social optimization case, the objective is to maximize the overall pay-off [20]. Naor shows that there is a threshold for the number of customers who

join the queue, both in the decentralized and centralized setting, and the threshold of the decentralized case is greater than the threshold of the centralized one. In other words, in the decentralized setting, individuals have negative externalities on others. In the observable part of our study, there are two parallel queues for public and private centres. Balking is not allowed, there is no reward (we assume that the reward of obtaining the service is equal for both centres, and also for all patients.), and there is a fee for joining the private centre. We extend Naor's results to a two dimensional queueing model, where we have two parallel servers. We prove that in the observable decentralized setting, for a given number of patients in the private/public centre, we have a threshold for the number of patients in the public/private centre. The same result is established for the observable centralized setting, by the use of MDP models. The public thresholds for the decentralized case is greater than or equal to the centralized one, and the private thresholds are smaller than or equal to the centralized setting.

There are papers which consider a queueing model which consists of a common queue served by two exponential servers. A related work is [21]. The arrivals are Poisson, and the service rates are different for each server. The length of the queue is observable. It is shown that an optimal policy exists which minimizes the average sojourn time of individuals in the system, and it is of a threshold-type. The individuals should be sent to the faster server whenever it is available, while they should be directed to the slower server if and only if the queue length exceeds a certain threshold value.

Another part of our work, the unobservable case, is related to the work of Burnetas and Georgiou [22] who analyse an unobservable system of two parallel Markovian single server queues with positive rewards and possibility of balking without any cost. Each customer decides whether to enter the system, and which of the queues to join to maximize her own pay-off. They formulate the system as a symmetric game played by all customers and show that a unique Nash equilibrium strategy exists, and they give an explicit expression for that. They also analyse the social optimization case in which the objective is to maximize the pay-off for the whole system. They show by a

numerical example that the proportion of the customers entering the system is lower in the socially optimal strategy.

In their model, they don't consider any cost for balking and this assumption makes their results different from ours. If we assume that all individuals who obtain the service have the same reward, and we also assume a high cost for balking, still the results are different. In the case that the reward of receiving service from the two centres are equal with a large value which prevents individual to balk from the system, the results in [22] and our results would be similar. In the unobservable case of our study, we also prove the existence of a unique symmetric Nash equilibrium in the decentralized setting, and a unique optimal strategy for the centralized setting.

Another relevant work for unobservable systems is Knight and Harper's study [23] who use routing games to calculate and compare the optimal solution and Nash equilibrium of a system of multiple unobservable queues with a single server where balking is allowed with a specific cost. They analyse the effect of allowing individuals to choose the service producer individually, by calculating the price of anarchy. It is shown that as the worth of getting service increases, the price of anarchy increases up to a point. They also show that the price of anarchy is low in an already inefficient system, so in these systems, letting individuals choose their servers is not considerably different from the social optimal solution.

The unobservable section of our work examines a specific model in which we deal with two M/M/1 models. So we analyse one special case of [23] in more detail. We find the structure of strategies in both decentralized and centralized cases.

Another paper which studies a system of two parallel M/M/1 servers is the work of Gue et al. [24]. They examine a self-financing two-tier queueing system, containing a free and a toll service. Arrivals are Poisson and the service time for both servers are exponentially distributed. Balking and reneging are not allowed. They assume that the capacity of the free server is fixed, and the capacity of toll server has to be built. Customers are identical, with the same waiting cost per unit of time. The length of the queue is unobservable, so customers have to choose between the free service and the

toll service based on the fee for joining the toll service and their expected waiting cost in each of the queues. They solve the system for the optimal capacity and fee of the toll service. They also characterize the solution's properties. As free-service capacity increases, the expected waiting time for both free and toll service also increases, which leads to a lower welfare. This result is analogous to the Downs-Thomson paradox in transportation economics.

This study is relevant to the unobservable part of our work with a finite buffer. In our study, all system parameters such as the buffer sizes of the queues, fee, waiting cost, and the arrival and service rates are fixed. We also consider a blocking cost, when the buffer is full. The individual's strategy in equilibrium is analysed, then we compare it with the social optimal strategy. We show by a numerical example that if we increase the buffer size of the free server, which is the public centre in our case, the total expected cost rate increases for the unobservable decentralized setting. We also show that the total expected cost for the unobservable centralized setting is decreasing in the buffer size of the free centre.

Another paper which considers a paradox of congestion in a queueing network is [25]. It is shown that in a congested queueing network, increasing the capacity would increase the mean transit time for individuals.

There are also some papers which compare the results of the two cases of observable and unobservable queues, or analyse the effect of getting more information about the number of individuals in the system. A related paper of this type with a game-theoretical point of view is by Hassin and Roet-Green [26] in which they analyse an unobservable single server queue with Poisson arrivals and exponential service time. Customers have three choices of joining the queue, balking, or inspecting the queue length and then decide to join or not. They consider the waiting cost, the cost of inspecting the queue length, and also the revenue of receiving service. Once a customer knows the queue length, she would obey the threshold strategy as in Naor's observable queue. Using the topological properties of the set of strategies in this game, they prove the existence and uniqueness of equilibrium in a two dimensional strategic game.

They conduct a sensitivity analysis on the preference of the customers to inspect the queue, and also on their willingness to join the queue after inspecting with respect to the revenue and cost of getting information. They also show that the ratio of waiting cost and the rate of service, leads to nonmonotonic behaviour. When the ratio is low, customers are more willing to join the queue without inspecting, when it is high, they tend to balk, and in between, customers would prefer to inspect the queue.

Many studies have been done related to routing policies in queueing systems. For example, in [27], they suppose two similar exponential servers with one arrival stream and only one queue. They show that in the observable case, in which the queue length at both servers are known, the optimal strategy is to route individuals to the shorter queue. While in the unobservable case, provided that the initial distribution of the two queue size is the same, the optimal strategy is to alternate between queues. These optimal strategies are independent of the statistics of the job arrival.

There are also some papers which work on the trade-off between the speed of service and its quality in a queueing system. Anand et al. [28] study a queueing system in which customers can choose whether or not to join a queue based on their self-interest. Considering the dependency between the service quality and service duration, they analyse the equilibrium behaviour of customers, the service rate, and also the pricing decisions. They have argued that the results from traditional queueing models are not applicable to their model when there is a relationship between the service quality and the service duration.

2.4 Scheduling Models in Healthcare

Many papers work on scheduling models in healthcare. Several studies consider the probability of "no-shows", in which a patient would cancel the appointment in the last minutes, or just doesn't show up. Green et al. [29] study the appointment system as a single server queueing model. They demonstrate that considering the"no-show" probability in a model has a significant impact on the performance of the system.

There are many papers which study different classes of customers arriving to the

system. Among those is the work of Xu et al. [30] which analyse the system of two stations with two parallel servers in each station. They assume two classes of customers with mutually independent Poisson arrivals. Class-1 customers have to receive service from station 1, but class-2 customers can be served by any of the stations. The service time is exponentially distributed with a common rate. The holding cost per unit of time for class-1 customers is assumed greater than or equal to the holding cost for class-2 customers. The objective is to minimize the expected discounted (or the long run average) holding cost by dynamically assigning customers to idle servers. The optimal policy is to assign a class-j customer to an idle server in station j, whenever possible, and a class-2 customer should be assigned to an idle server in station 1 only when there is no class-1 customer waiting, and the number of customers waiting in queue 2 exceeds a threshold. Furthermore, this threshold increases in the number of busy servers in station 1.

Another relevant paper which considers two classes of customers is by Suk and Cassandras [31]. They study the dynamic scheduling problem for a queueing system with two classes of customers with finite queue capacity, competing for service at a single station. The cost is assumed to be linear in the number of customers in the queues, and there is also a blocking cost for each queue. They show that if the blocking cost of each queue is greater than or equal to the holding cost, then the optimal policy is characterized by a switching curve. If for one of the queues, the blocking cost is less than the holding cost, and for the other queue it is greater than or equal to the holding cost, a fixed priority rule is optimal. Suk and Cassandras examine an extreme case, when the blocking cost of one queue is sufficiently larger than the other, whereas the holding cost of that queue is smaller than the other. In this case, the optimal policy is of a threshold-type policy.

A more general case was studied by Baras et al. [32], in which the optimality of μc-rule was shown for expected long run average and expected discounted criterion, over both finite and infinite horizons. A more recent study on an M/M/c queue with two priority classes is [33].

Chapter 3

MODEL

This study analyses a healthcare system with two parallel centres, one public and one private. The two centres offer different service rates at different fees. Specifically, the private centre offers a faster service as compared to the public one and charges a fee $f > 0$ while the public centre is free. Each centre is assumed to operate with a single server and exponentially distributed service time. The service rates for the public and private centres are denoted by $\mu_1 > 0$ and $\mu_2 > 0$, respectively. Patients join the system according to a Poisson process with rate $\lambda > 0$. The buffer sizes for both queues are infinite. An arriving patient either joins the public-centre queue or the private-center queue so balking is not allowed. Moreover a patient who joins a queue does not leave it until her service is completed, i.e., reneging is not allowed. For each patient, the waiting cost per time unit spent in the system, whether waiting or being served, is $c > 0$. We assume that the reward of receiving service from both centres is the same for all patients. As it doesn't have any effect on individuals' choice, we only consider the costs. The parameters are known to everyone. The main assumptions on the parameters are:

(A1) $\mu_1 < \mu_2$,

(A2) $\lambda < \mu_1 + \mu_2$.

(A1) implies that the private server is faster. Practically, in some healthcare centres, this assumption may not hold. Sometimes private centres perform more tests and the doctors in these centres spend more time for each patient rather than the public centres. In this work, we focus on the diagnosis facilities and assume that in the private centre, tests can be done faster. Assumption (A2) prevents having a queue

with an infinite length. The model is shown in Figure 3.1.

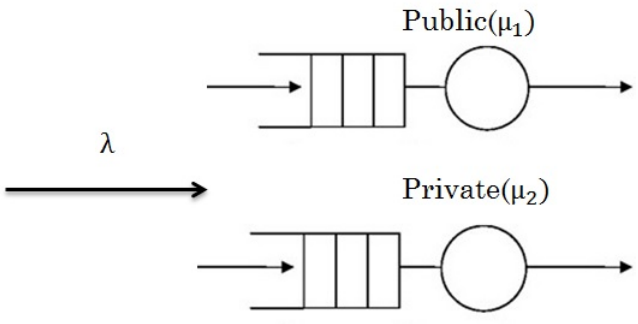

Figure 3.1: Model

We analyse this model in decentralized and centralized settings, and also in observable and unobservable cases.

In the observable decentralized setting with infinite buffer, individuals observe the length of the queues and then choose between joining the public and the private centre in order to minimize their own expected cost. In the observable centralized setting, there is a central authority who observes the number of individuals in each centre and sends the arriving patients to either of the public or the private centres to minimize the expected cost of the whole system.

In the unobservable setting with infinite buffer, the objective is to find the optimal p, which is the probability of joining the public centre. As balking is not allowed, the probability of joining the private centre is $1 - p$. In the unobservable decentralized setting, individuals act in order to minimize their own expected cost in the symmetric simultaneous-move game. We prove the existence of a unique symmetric Nash equilibrium, and derive explicit expressions for finding p. In the unobservable centralized setting, a central authority decides on the optimal p in order to minimize the expected cost of the whole system. In this case, we also find explicit expressions to find the optimal p.

To extend our model to the system with finite buffer, we consider a blocking cost (c_B)

for each individual who encounters a full system. The buffer sizes for the public and the private centres are denoted by m_1 and m_2, respectively. We extend our results in the observable case to the model with finite buffer. In the unobservable case, we show via a numerical example that our results hold for the model with finite buffer under some conditions.

Chapter 4

OBSERVABLE CASE

This section analyses the situation where the number of patients in each centre can be observed by the decision maker at any point in time.

4.1 Decentralized Setting with Infinite Buffer

In the decentralized setting, the state of the system is the number of patients waiting in each centre, which we denote by (n_1, n_2) with integer $n_i \geq 0$ for $i = 1, 2$. In this case, arriving patients choose which centre to join with the objective of minimizing their individual expected costs. An arriving patient that observes n_1 patients in the public centre and n_2 patients in the private centre (both including those in service) incurs an expected cost of

$$C_1(n_1) \equiv \frac{c(n_1 + 1)}{\mu_1}, \tag{4.1}$$

if she joins the public centre, and

$$C_2(n_2) \equiv f + \frac{c(n_2 + 1)}{\mu_2}, \tag{4.2}$$

if she joins the private center.

To minimize her expected cost, she will choose the public centre if $C_1(n_1) \leq C_2(n_2)$, and the private centre if $C_1(n_1) > C_2(n_2)$. This particular strategy assumes that the patient chooses the public centre if indifferent. Other optimal strategies are those that assign different probabilities to joining the public centre in case of indifference.

A strategy is a Nash equilibrium strategy if and only if no player could increase her expected pay-off by deviating from that strategy [34]. Let $p_o^e(n_1, n_2)$ be the equilibrium probability of joining to the public centre when there are n_1 patients in the public

centre and n_2 patients in the private centre. The equilibrium strategy for this patient is

$$p_o^e(n_1, n_2) = \begin{cases} 1 & \text{if } n_1 \leq (n_2+1)\frac{\mu_1}{\mu_2} + \frac{f}{c}\mu_1 - 1, \\ 0 & \text{if } n_1 > (n_2+1)\frac{\mu_1}{\mu_2} + \frac{f}{c}\mu_1 - 1. \end{cases} \tag{4.3}$$

This implies that for any given n_2, there is a threshold $n_1(n_2)$ such that the patient joins the public centre if $n_1 \leq n_1(n_2)$ and the private centre otherwise. Analogously, for any given n_1, there is a threshold $n_2(n_1)$ such that the patient joins the public centre if $n_2 \geq n_2(n_1)$ and the private centre otherwise. The following theorem states this result and defines the thresholds.

Theorem 1. *In the observable decentralized setting with infinite buffer, an arriving patient that observes n_1 and n_2 patients in the public and private centres, respectively, joins the public centre if $n_1 \leq n_1(n_2)$ and the private centre otherwise, or equivalently the public centre if $n_2 \geq n_2(n_1)$ and the private centre otherwise. The integer thresholds $n_1(n_2)$ and $n_2(n_1)$ are given by:*

$$n_1(n_2) = \left\lfloor (n_2+1)\frac{\mu_1}{\mu_2} + \frac{f}{c}\mu_1 - 1 \right\rfloor, \tag{4.4}$$

$$n_2(n_1) = \left\lfloor (n_1+1)\frac{\mu_2}{\mu_1} - \frac{f}{c}\mu_2 \right\rfloor. \tag{4.5}$$

Consequently, the equilibrium strategy (4.3) can then be expressed as a threshold strategy:

$$p_o^e(n_1, n_2) = \begin{cases} 1 & \text{if } n_1 \leq n_1(n_2), \\ 0 & \text{if } n_1 > n_1(n_2), \end{cases} \tag{4.6}$$

or equivalently,

$$p_o^e(n_1, n_2) = \begin{cases} 1 & \text{if } n_2 \geq n_2(n_1), \\ 0 & \text{if } n_2 < n_2(n_1). \end{cases} \tag{4.7}$$

Corollary 1. *(i) For fixed n_2, $n_1(n_2)$ is independent of λ, nonincreasing in c and μ_2, and nondecreasing in f and μ_1.*
(ii) For fixed n_1, $n_2(n_1)$ is independent of λ, nonincreasing in f and μ_1, and nondecreasing in c and μ_2.

Proof. To prove that for fixed n_1, $n_2(n_1)$ is nondecreasing in μ_2, we can write equation (4.5) as

$$n_2(n_1) = \left\lfloor \left((n_1+1)\frac{1}{\mu_1} - \frac{f}{c} \right) \mu_2 \right\rfloor$$

For $(n_1+1)\frac{1}{\mu_1} - \frac{f}{c} \geq 0$, $n_2(n_1)$ is nondecreasing in μ_2, and for $(n_1+1)\frac{1}{\mu_1} - \frac{f}{c} < 0$, $n_2(n_1)$ is defined as 0, so the proof is complete. Other results of Corollary 1 can be deduced directly from (4.4) and (4.5). $\square$

The purpose of the rest of this section is to understand the dynamics of the observable decentralized system where each arriving patient adopts the equilibrium strategy p_o^e. To do this, let

$$n_1^k \equiv \begin{cases} n_1(0)+1 & \text{for } k=1, \\ n_1\left(n_2^{k-1}\right)+1 & \text{for } k=2,3,\ldots, \end{cases} \tag{4.8}$$

and

$$n_2^k \equiv n_2(n_1^k) \quad \text{for } k=1,2,3,\ldots. \tag{4.9}$$

The numbers n_1^k and n_2^k stand for the *switching levels* for the public and private centres respectively. The transition rates between the states change at these points, so for calculation purposes, these points require special attention when setting the steady-state equations of the system. To illustrate this phenomenon, note that all arriving patients go to the public centre on the set of states $\{(n_1,n_2) : n_1 \leq n_1^1 - 1\}$ and to the private centre on the set of states $\{(n_1,n_2) : n_1 \geq n_1^1, n_2 < n_2^1\}$. Figure 4.1 exhibits this structure in a general form.

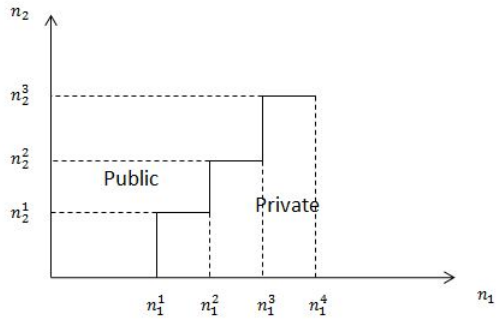

Figure 4.1: Optimal Policy in Observable Decentralized Setting

The following theorem provides an explicit representation of switching levels.

Theorem 2. *For any $k \in \{1, 2, \dots\}$, the switching levels n_i^k satisfy:*

$$n_1^k = k - 1 + \left\lfloor \frac{f}{c}\mu_1 + \frac{\mu_1}{\mu_2} \right\rfloor, \tag{4.10}$$

$$n_2^k = \left\lfloor \left(k + \left\lfloor \frac{f}{c}\mu_1 + \frac{\mu_1}{\mu_2} \right\rfloor \right) \frac{\mu_2}{\mu_1} - \frac{f}{c}\mu_2 \right\rfloor. \tag{4.11}$$

Proof. First note that if (4.10) holds, then by (4.5) and (4.9), we have:

$$n_2^k = n_2(n_1^k) = \left\lfloor \left(k + \left\lfloor \frac{f}{c}\mu_1 + \frac{\mu_1}{\mu_2} \right\rfloor \right) \frac{\mu_2}{\mu_1} - \frac{f}{c}\mu_2 \right\rfloor,$$

so (4.10) implies (4.11). To show (4.10), first note that for $k = 1$, by definition (4.8), n_1^1 can be written as follows:

$$n_1^1 = n_1(0) + 1 = \left\lfloor \frac{\mu_1}{\mu_2} + \frac{f}{c}\mu_1 - 1 \right\rfloor + 1 = k - 1 + \left\lfloor \frac{f}{c}\mu_1 + \frac{\mu_1}{\mu_2} \right\rfloor,$$

Hence (4.10) is true for $k = 1$, which starts the induction. Suppose (4.10) holds for

some $k \geq 1$, which is the induction hypothesis, implying that (4.11) holds for k. Then

$$\begin{aligned} n_1^{k+1} &= n_1\left(n_2^k\right)+1 = \left\lfloor \left\{ \left\lfloor \left(k + \left\lfloor \frac{f}{c}\mu_1 + \frac{\mu_1}{\mu_2} \right\rfloor \right) \frac{\mu_2}{\mu_1} - \frac{f}{c}\mu_2 \right\rfloor + 1 \right\} \frac{\mu_1}{\mu_2} + \frac{f}{c}\mu_1 - 1 \right\rfloor + 1 \\ &= \left\lfloor \left\{ \left\lfloor \left(k + \left\lfloor \frac{f}{c}\mu_1 + \frac{\mu_1}{\mu_2} \right\rfloor \right) \frac{\mu_2}{\mu_1} - \frac{f}{c}\mu_2 \right\rfloor + 1 \right\} \frac{\mu_1}{\mu_2} + \frac{f}{c}\mu_1 \right\rfloor \\ &\leq \left\lfloor \left\{ \left(k + \left\lfloor \frac{f}{c}\mu_1 + \frac{\mu_1}{\mu_2} \right\rfloor \right) \frac{\mu_2}{\mu_1} - \frac{f}{c}\mu_2 + 1 \right\} \frac{\mu_1}{\mu_2} + \frac{f}{c}\mu_1 \right\rfloor \\ &= \left\lfloor k + \left\lfloor \frac{f}{c}\mu_1 + \frac{\mu_1}{\mu_2} \right\rfloor - \frac{f}{c}\mu_1 + \frac{\mu_1}{\mu_2} + \frac{f}{c}\mu_1 \right\rfloor = k + \left\lfloor \frac{f}{c}\mu_1 + \frac{\mu_1}{\mu_2} \right\rfloor, \end{aligned}$$

where the first equality follows from (4.8), the second from (4.4) and (4.11), and the rest from the properties of floor function and assumption $\mu_1 < \mu_2$. Similarly,

$$\begin{aligned} n_1^{k+1} &\geq \left\lfloor \left\{ \left(k + \left\lfloor \frac{f}{c}\mu_1 + \frac{\mu_1}{\mu_2} \right\rfloor \right) \frac{\mu_2}{\mu_1} - \frac{f}{c}\mu_2 - 1 + 1 \right\} \frac{\mu_1}{\mu_2} + \frac{f}{c}\mu_1 \right\rfloor \\ &= \left\lfloor k + \left\lfloor \frac{f}{c}\mu_1 + \frac{\mu_1}{\mu_2} \right\rfloor - \frac{f}{c}\mu_1 + \frac{f}{c}\mu_1 \right\rfloor = k + \left\lfloor \frac{f}{c}\mu_1 + \frac{\mu_1}{\mu_2} \right\rfloor. \end{aligned}$$

Hence $n_1^{k+1} = k + \left\lfloor \frac{f}{c}\mu_1 + \frac{\mu_1}{\mu_2} \right\rfloor$, which completes the proof. □

Theorem 2 allows calculating all switching levels by using the model parameters μ_1, μ_2, f, c. Corollary 2 describes the changes in switching levels as the model parameters change.

Corollary 2. *For $k = 1, 2, ...,$*

(i) n_1^k and n_2^k are independent of λ.

(ii) n_1^k is nondecreasing in f and μ_1, and nonincreasing in c and μ_2.

Proof. Parts $(i) - (ii)$ immediately follow from (4.10)-(4.11). □

The monotonicity results stated in Corollary 2 are intuitive, since one would expect the customers to prefer the public centre more if the private service is accompanied with a higher fee, if the waiting cost is lower, if the public centre offers service at a higher rate, or if the private centre has a lower service rate. The following corollary of Theorem 2 establishes the relations between n_1^k and n_1^{k+1}, n_2^k and n_2^{k+1}, n_1^k and n_2^k.

Corollary 3. *For $k \in 1, 2, ...,$ the switching levels n_i^k satisfy:*

(i) $n_1^{k+1} = n_1^k + 1,$

(ii) $n_2^k + \left\lfloor \frac{\mu_2}{\mu_1} \right\rfloor \leq n_2^{k+1} \leq n_2^k + \frac{\mu_2}{\mu_1} + 1,$

(iii) $1 + \left\lfloor \frac{(k-1)\mu_2}{\mu_1} \right\rfloor \leq n_2^k \leq \left\lfloor \frac{k\mu_2}{\mu_1} \right\rfloor + 1,$

(iv) If $n_2^k > n_1^k$ for some $k \geq 1$, then $n_2^{k+1} > n_1^{k+1}$.

Proof. Part (i) immediately follows from (4.10). To show (ii), we have:

$$n_2^{k+1} = \left\lfloor \left(k + \left\lfloor \frac{f}{c}\mu_1 + \frac{\mu_1}{\mu_2} \right\rfloor \right) \frac{\mu_2}{\mu_1} + \frac{\mu_2}{\mu_1} - \frac{f}{c}\mu_2 \right\rfloor \geq n_2^k + \left\lfloor \frac{\mu_2}{\mu_1} \right\rfloor,$$

and

$$n_2^{k+1} \leq \left(k + \left\lfloor \frac{f}{c}\mu_1 + \frac{\mu_1}{\mu_2} \right\rfloor \right) \frac{\mu_2}{\mu_1} + \frac{\mu_2}{\mu_1} - \frac{f}{c}\mu_2 \leq n_2^k + 1 + \frac{\mu_2}{\mu_1},$$

which proves (ii). To show (iii), we have:

$$n_2^k \leq \left\lfloor \left(k + \frac{f}{c}\mu_1 + \frac{\mu_1}{\mu_2} \right) \frac{\mu_2}{\mu_1} - \frac{f}{c}\mu_2 \right\rfloor = \left\lfloor \frac{k\mu_2}{\mu_1} \right\rfloor + 1,$$

$$n_2^k \geq \left\lfloor \left(k + \frac{f}{c}\mu_1 + \frac{\mu_1}{\mu_2} - 1 \right) \frac{\mu_2}{\mu_1} - \frac{f}{c}\mu_2 \right\rfloor = 1 + \left\lfloor \frac{(k-1)\mu_2}{\mu_1} \right\rfloor,$$

which completes the proof of (iii). Finally, since $n_1^{k+1} = n_1^k + 1$ and by (ii),

$$n_2^{k+1} \geq n_2^k + \left\lfloor \frac{\mu_2}{\mu_1} \right\rfloor \geq n_2^k + 1 > n_1^k + 1 = n_1^{k+1},$$

where the second inequality follows since $\mu_1 < \mu_2$ and the third inequality from $n_2^k > n_1^k$. $\square$

To assess the performance of the system in the long run, we need steady-state probabilities. If a finite stationary distribution exists, it has to satisfy the steady-state equations (see the Appendix). The following theorem proves the stability of the system under (A2), i.e., $\lambda < \mu_1 + \mu_2$.

Theorem 3. *In the observable setting, under (A2), the system is stable, so the queue size for both public and private centres are finite.*

Proof. By contradiction, suppose that the arrival rate to the public centre is greater than μ_1. Let the public arrival rate be $\mu_1 + \epsilon$, so the private arrival rate would be $\lambda - \mu_1 - \epsilon$. In this case, the expected waiting time in the public centre goes to infinity, while in the private centre, the expected waiting time is finite. As the objective is to minimize the expected cost, this case never happens, and individuals tend to join the queue which has a smaller waiting time, rather than the unstable queue. So the assumption does not hold and the system is stable. □

For queueing systems with two dimensional states, finding a closed-form expression for the steady-state probability distribution is usually difficult, unless the system have a product form solution. Resing and Örmeci encounter the same problem for a tandem queue with a shared server [35]. To find an approximation for these probability distribution, we consider systems which have a finite capacity room.

4.2 Decentralized Setting with Finite Buffer

Let m_1 and m_2 be the buffer sizes for the public and private centres, respectively. The introduction of buffer size into the model affects only the switching levels at the boundaries, when at least one of the buffers is full.

Theorem 4. *In the observable decentralized setting with finite buffer, an arriving patient that observes n_1 and n_2 patients in the public and private centres, respectively, joins the public centre if $n_1 \leq n_1(n_2)$ and the private centre otherwise, or equivalently the public centre if $n_2 \geq n_2(n_1)$ and the private centre otherwise. The integer thresholds $n_1(n_2)$ and $n_2(n_1)$ are*

$$n_1(n_2) \equiv \begin{cases} \min\left\{\left\lfloor (n_2+1)\frac{\mu_1}{\mu_2} + \frac{f}{c}\mu_1 - 1 \right\rfloor, m_1 - 1\right\} & \text{for } n_2 \neq m_2, \\ m_1 - 1 & \text{for } n_2 = m_2, \end{cases} \tag{4.12}$$

and

$$n_2(n_1) \equiv \begin{cases} \min\left\{\left\lfloor (n_1+1)\frac{\mu_2}{\mu_1} - \frac{f}{c}\mu_2 \right\rfloor, m_2\right\} & \text{for } n_1 \neq m_1, \\ m_2 & \text{for } n_1 = m_1. \end{cases} \tag{4.13}$$

The steady-state equations for this case are the same as the model with infinite buffer, except at the boundaries (see the Appendix).

Figure 4.2 shows the structure of switching levels on a numerical example with finite buffer. In this example, the switching levels are 4, 5, and 6 for the public centre and 1, 3, and 4 for the private centre.

4.3 Centralized Setting with Infinite Buffer

In the centralized setting, there is an authority that makes the decision for every patient with the objective of minimizing the total discounted cost of all patients, i.e., the sum of waiting costs and the fees for joining the private centre. Under the observability assumption, the authority can choose a policy that depends on the number of individuals in each centre. The system can be modelled as a Markov decision process. We let the state of the system be (n_1, n_2) with integer $n_i \geq 0$ for $i = 1, 2$. $\tilde{C}(n_1, n_2)$ is the minimum total discounted cost of the system when n_1 and n_2 patients are in the public and private centres, respectively, over an infinite horizon. To obtain a discrete-time model, we assume (without loss of generality) that $\lambda + \mu_1 + \mu_2 + \beta = 1$, where β is a positive scalar, and can be interpreted as the rate of the system's failure.

For $n_1 \geq 0$ and $n_2 \geq 0$, the optimality equation is

$$\begin{aligned} \tilde{C}(n_1, n_2) &= \lambda \min\left\{\tilde{C}(n_1+1, n_2), f + \tilde{C}(n_1, n_2+1)\right\} \\ &+ \mu_1 \tilde{C}\left((n_1-1)^+, n_2\right) + \mu_2 \tilde{C}\left(n_1, (n_2-1)^+\right) \\ &+ c(n_1+n_2) \end{aligned} \tag{4.14}$$

This model was analysed in [36] when the waiting costs in the two centres are different and there is no fee (f). Here, we have the same waiting cost for both centres, and a positive fee for joining the private queue. We employ value iteration to prove the structure of the optimal policy. To that end, let $\tilde{C}_t\left(n_1, n_2\right)$ be the minimum expected discounted cost when there are n_1 patients in the public and n_2 patients in the private centre, and there are t transition epochs remaining in the horizon, with $\tilde{C}_0(n_1, n_2) = 0$. For $n_1 \geq 0$ and $n_2 \geq 0$

$$\begin{aligned} \tilde{C}_t\left(n_1, n_2\right) &= \lambda \min\left\{\tilde{C}_{t-1}\left(n_1+1, n_2\right), f + \tilde{C}_{t-1}\left(n_1, n_2+1\right)\right\} \\ &+ \mu_1 \tilde{C}_{t-1}\left((n_1-1)^+, n_2\right) + \mu_2 \tilde{C}_{t-1}\left(n_1, (n_2-1)^+\right) \\ &+ c(n_1+n_2) \end{aligned} \tag{4.15}$$

[36] shows that $\tilde{C}_t$ converges to $\tilde{C}$ as $t \longrightarrow \infty$.

The difference between the expected discounted cost of joining the public centre, and the private centre in state (n_1, n_2) is defined as:

$$\Delta(n_1, n_2) = \tilde{C}(n_1+1, n_2) - \tilde{C}(n_1, n_2+1) - f$$

Lemma 1. *$\Delta(n_1, n_2)$ is nondecreasing in n_1 for each fixed n_2, and nondecreasing in n_2 for each fixed n_1.*

Proof. We prove that $\Delta_t(n_1, n_2)$ is nondecreasing in n_1 for each n_2 and t by induction. The proof is the same for the monotonicity in n_2.

The claim holds for $t = 0$ since $\Delta_0(n_1, n_2) = 0$. Suppose it holds for $t \geq 0$, then

$$\begin{aligned}
\Delta_{t+1}(n_1, n_2) &= \tilde{C}_{t+1}(n_1 + 1, n_2) - \tilde{C}_{t+1}(n_1, n_2 + 1) \\
&= \lambda \min\left\{\tilde{C}_t(n_1 + 2, n_2), f + \tilde{C}_t(n_1 + 1, n_2 + 1)\right\} \\
&+ \mu_1 \tilde{C}_t(n_1, n_2) + \mu_2 \tilde{C}_t(n_1 + 1, (n_2 - 1)^+) + c(n_1 + n_2 + 1) \\
&- \lambda \min\left\{\tilde{C}_t(n_1 + 1, n_2 + 1), f + \tilde{C}_t(n_1, n_2 + 2)\right\} \\
&+ \mu_1 \tilde{C}_t((n_1 - 1)^+, n_2 + 1) + \mu_2 \tilde{C}_t(n_1, n_2) + c(n_1 + n_2 + 1) \\
&= \mu_1 \left[\tilde{C}_t(n_1, n_2) - \tilde{C}_t((n_1 - 1)^+, n_2 + 1)\right] \\
&+ \mu_2 \left[\tilde{C}_t(n_1 + 1, (n_2 - 1)^+) - \tilde{C}_t(n_1, n_2)\right] \\
&+ \lambda \left[\min\left\{\tilde{C}_t(n_1 + 2, n_2), f + \tilde{C}_t(n_1 + 1, n_2 + 1)\right\}\right] \\
&- \lambda \left[\min\left\{\tilde{C}_t(n_1 + 1, n_2 + 1), f + \tilde{C}_t(n_1, n_2 + 2)\right\}\right]
\end{aligned}$$

From the induction hypothesis the first and the second terms are monotonically nondecreasing in n_1. The remaining terms can be written as

$$\begin{aligned}
& \lambda \left[\tilde{C}_t(n_1 + 1, n_2 + 1) + f + \min\left\{\tilde{C}_t(n_1 + 2, n_2) - \tilde{C}_t(n_1 + 1, n_2 + 1) - f, 0\right\}\right] \\
- \; & \lambda \left[\tilde{C}_t(n_1 + 1, n_2 + 1) + \min\left\{0, f + \tilde{C}(n_1, n_2 + 2) - \tilde{C}_t(n_1 + 1, n_2 + 1)\right\}\right] \\
= \; & \lambda \left[\min\left\{\tilde{C}_t(n_1 + 2, n_2) - \tilde{C}_t(n_1 + 1, n_2 + 1) - f, 0\right\}\right] \\
+ \; & \lambda \left[\max\left\{0, \tilde{C}_t(n_1 + 1, n_2 + 1) - \tilde{C}(n_1, n_2 + 2) - f\right\}\right] + \lambda f \\
= \; & \lambda \left[\min\left\{\Delta_1^t(n_1 + 1, n_2), 0\right\} + \max\left\{0, \Delta_1^t(n_1, n_2 + 1)\right\} + f\right]
\end{aligned}$$

Since $\Delta_t(n_1 + 1, n_2)$ and $\Delta_t(n_1, n_2 + 1)$ are nondecreasing in n_1 by the induction hypothesis, the same is true for the preceding expression. Since $\Delta_{t+1}(n_1, n_2)$ converges to $\Delta(n_1, n_2)$, $\Delta(n_1, n_2)$ must be nondecreasing in n_1. $\square$

The following theorem proves the optimality of a threshold type policy.

Theorem 5. *In the observable centralized setting with infinite buffer,*
(i) Given n_2, there is a threshold $\tilde{n}_1(n_2)$ such that it is optimal to assign an arriving patient to public centre if $n_1 \leq \tilde{n}_1(n_2)$ and to the private centre otherwise.

(ii) Given n_1, there is a threshold $\tilde{n}_2(n_1)$ such that it is optimal to assign an arriving patient to the private centre if $n_2 \leq \tilde{n}_2(n_1)$ and to public centre otherwise.

Proof. It follows from Lemma 1. □

In the observable centralized case, we also define switching levels as the points where the transition rates between the states change. These points are useful to derive the steady-state equations. $\tilde{n}_1^k$ and $\tilde{n}_2^k$ are the switching levels for the public and the private centre, respectively. In the following corollary, the relationship between $\tilde{n}_1^k$ and $\tilde{n}_1^{k+1}$ is shown.

Corollary 4. *The differences between two consequent switching levels of the public centre is always 1. Accordingly, $\tilde{n}_1^{k+1} = \tilde{n}_1^k + 1$.*

Proof. We prove that if the authority sends the patient to private centre in state (n_1, n_2), it also sends her to private centre in state $(n_1 + 1, n_2 + 1)$.

$$\tilde{C}(n_1 + 1, n_2) > \tilde{C}(n_1, n_2 + 1) + f \Rightarrow \tilde{C}(n_1 + 2, n_2 + 1) > \tilde{C}(n_1 + 1, n_2 + 2) + f$$

By Lemma 1, $\Delta(n_1, n_2)$ is nondecreasing in both n_1 and n_2, so

$$\tilde{C}(n_1 + 2, n_2 + 1) - \tilde{C}(n_1 + 1, n_2 + 2) - f \geq \tilde{C}(n_1 + 1, n_2) - \tilde{C}(n_1, n_2 + 1) - f$$

which is a sufficient condition for the claim, and the proof is complete. □

To explore the effects of different parameters such as c, f, λ, μ_1, and μ_2, we introduce $\Delta_t^{\gamma}(n_1, n_2)$ which is the difference in the expected discounted cost of joining the public and the private centres in state (n_1, n_2), with respect to the parameter γ.

$$\Delta_{t+1}^{\gamma}(n_1, n_2) = \tilde{C}_{t+1}(n_1 + 1, n_2, \gamma) - \tilde{C}_{t+1}(n_1, n_2 + 1, \gamma)$$

Lemma 2. *Given (n_1, n_2), $\Delta_t^{\gamma}(n_1, n_2)$ is nonincreasing in $\gamma = c, \mu_2$ and nondecreasing in $\gamma = f, \lambda, \mu_1$.*

Proof. We prove that Δ_t^γ is monotone in γ by induction. It is true for $t = 0$. Suppose it is true for $t \geq 0$ and consider

$$\begin{aligned}
\Delta_{t+1}^\gamma &= \tilde{C}_{t+1}(n_1+1, n_2, \gamma) - \tilde{C}_{t+1}(n_1, n_2+1, \gamma) \\
&= \lambda \min\left\{\tilde{C}_t(n_1+2, n_2, \gamma), \tilde{C}_t(n_1+1, n_2+1, \gamma) + f\right\} \\
&+ \mu_1 \tilde{C}_t(n_1, n_2, \gamma) + \mu_2 \tilde{C}_t\left(n_1+1, (n_2-1)^+, \gamma\right) + c(n_1+n_2+1) \\
&- \lambda \min\left\{\tilde{C}_t(n_1+1, n_2+1, \gamma), \tilde{C}_t(n_1, n_2+2, \gamma) + f\right\} \\
&- \mu_1 \tilde{C}_t\left((n_1-1)^+, n_2+1, \gamma\right) - \mu_2 \tilde{C}_t(n_1, n_2, \gamma) - c(n_1+n_2+1) \\
&= \mu_1 \left[\tilde{C}_t(n_1, n_2, \gamma) - \tilde{C}_t\left((n_1-1)^+, n_2, \gamma\right)\right] \\
&+ \mu_2 \left[\tilde{C}_t\left(n_1+1, (n_2-1)^+, \gamma\right) - \tilde{C}_t(n_1, n_2, \gamma)\right] \\
&+ \lambda \min\left\{\tilde{C}_t(n_1+2, n_2, \gamma), \tilde{C}_t(n_1+1, n_2+1, \gamma) + f\right\} \\
&- \lambda \min\left\{\tilde{C}_t(n_1+1, n_2+1, \gamma), \tilde{C}_t(n_1, n_2+2, \gamma) + f\right\}
\end{aligned}$$

From the induction hypothesis the first and second terms are monotone in γ. The remaining can be written as:

$$\begin{aligned}
& \lambda \left[\min\left\{\tilde{C}_t(n_1+2, n_2, \gamma) - \tilde{C}_t(n_1+1, n_2+1, \gamma) - f, 0\right\} + \tilde{C}_t(n_1+1, n_2+1, \gamma) + f\right] \\
- \; & \lambda \left[\min\left\{0, \tilde{C}_t(n_1, n_2+2, \gamma) - \tilde{C}_t(n_1+1, n_2+1, \gamma) + f\right\} + \tilde{C}_t(n_1+1, n_2+1, \gamma)\right] \\
= \; & \lambda \left[\min\left\{\tilde{C}_t(n_1+2, n_2, \gamma) - \tilde{C}_t(n_1+1, n_2+1, \gamma) - f, 0\right\} + f\right] \\
+ \; & \lambda \left[\max\left\{\tilde{C}_t(n_1+1, n_2+1, \gamma) - \tilde{C}_t(n_1, n_2+2, \gamma) - f, 0\right\}\right]
\end{aligned}$$

Since $\Delta_t^\gamma(n_1+1, n_2)$ and $\Delta_t^\gamma(n_1, n_2+1)$ are monotone in γ by the induction hypothesis, the same is true for the preceding expression. As each term of $\Delta_{t+1}^\gamma(n_1, n_2)$ is monotone in (n_1, n_2), the induction proof is complete. □

The effect of changes in each parameter of our model to the observable centralized thresholds are shown in the next corollary.

Corollary 5. *The public switching level $\tilde{n}_1^k$ is nondecreasing in f, μ_1, and λ, and nonincreasing in c and μ_2.*

The private switching level $\tilde{n}_2^k$ is nondecreasing in c, μ_2, and λ, and nonincreasing in f and μ_1.

Proof. As the public and private switching levels are dependant of $\Delta(n_1, n_2)$, the results in this corollary can be deduced by Lemma 2. □

The thresholds and switching levels in the observable decentralized setting are independent of λ (by Corollary 1), while for observable centralized setting they are nondecreasing in λ (by Corollary 5). This result is intuitive, as in the decentralized case, the arrival rate does not affect the decision of individuals. The patients decide which queue to join with the objective of minimizing their own expected cost. However, in the centralized setting, the central authority sends individuals to either of the public or the private centres to minimize the total expected cost of the system.

Long-run Average Cost

Another way to analyse the dynamic model is by using the long-run average costs. In our model, the Markov chain is a uni-chain because under stability conditions, $(0, 0)$ is always reachable from all states. By [37], the long-run average costs converge to the minimum expected discounted costs of the system.

4.4 Centralized Setting with Finite Buffer

Suppose that the buffer size for the public centre is m_1, and it is m_2 for the private centre. The blocking cost is c_B, which occurs when the system is full.

The optimality equations for the model with finite buffer

$$\begin{aligned}
\tilde{C}(n_1,n_2) &= \lambda \min\left\{\tilde{C}(n_1+1,n_2), f+\tilde{C}(n_1,n_2+1)\right\} && (4.16)\\
&+ \mu_1\tilde{C}\left((n_1-1)^+,n_2\right)+\mu_2\tilde{C}\left(n_1,(n_2-1)^+\right)\\
&+ c(n_1+n_2) \quad \text{for } 0\le n_1<m_1 \text{ and } 0\le n_2<m_2\\
\tilde{C}(m_1,n_2) &= \lambda(f+\tilde{C}(m_1,n_2+1)) && (4.17)\\
&+ \mu_1\tilde{C}(m_1-1,n_2)+\mu_2\tilde{C}\left(m_1,(n_2-1)^+\right)\\
&+ c(m_1+n_2) \quad \text{for } 0\le n_2<m_2\\
\tilde{C}(n_1,m_2) &= \lambda\tilde{C}(n_1+1,m_2)) && (4.18)\\
&+ \mu_1\tilde{C}\left((n_1-1)^+,m_2\right)+\mu_2\tilde{C}(n_1,m_2-1)\\
&+ c(n_1+m_2) \quad \text{for } 0\le n_1<m_1\\
\tilde{C}(m_1,m_2) &= \lambda c_B && (4.19)\\
&+ \mu_1\tilde{C}(m_1-1,m_2)+\mu_2\tilde{C}(m_1,m_2-1)+c(m_1+m_2)
\end{aligned}$$

The following theorem states that the model with finite buffer also has a threshold-type optimal policy.

Theorem 6. *In the observable centralized setting with finite buffer, there is a threshold type optimal policy.*
(i) Given n_2, there is a threshold $\tilde{n}_1(n_2)$ such that it is optimal to assign an arriving patient to public centre if $n_1 \le \tilde{n}_1(n_2)$ and to private centre else.
(ii) Given n_1, there is a threshold $\tilde{n}_2(n_1)$ such that it is optimal to assign an arriving patient to private centre if $n_2 \le \tilde{n}_2(n_1)$ and to public centre otherwise.

Proof. The proof is similar to the case with infinite buffer. To complete the proof, we only need to check the boundaries, which are the states (m_1, n_2) and (n_1, m_2). According to Theorem (5), if joining the public centre is the optimal strategy in state (n_1, n_2), it is also the optimal strategy in state $(n_1 - 1, n_2)$. This is always true for state (n_1, m_2) where the private centre is full and the patient is sent to the public centre. Equivalently, this theorem implies that if joining the private centre is the

optimal strategy in state $(n_1 - 1, n_2)$, it is also the optimal strategy in state (n_1, n_2). In (m_1, n_2), as the public centre is full, the individual is sent to the private centre. Theorem (5) also shows that if the optimal strategy in state (n_1, n_2) is joining the private centre, it is also the optimal strategy in state $(n_1, n_2 - 1)$. This is always true for state (m_1, n_2) where the public centre is full and the patient is sent to the private centre. Equivalently, this theorem implies that if the optimal strategy in state $(n_1, n_2 - 1)$ is to join the public centre, it is also the optimal strategy in state (n_1, n_2). In (n_1, m_2), as the private centre is full, the individual is sent to public centre. □

In the numerical examples, we apply relative value iteration to find the average cost in the centralized setting. The thresholds for a numerical example are shown in Figure 4.3.

4.5 Centralized vs. Decentralized Setting

In this section, we derive a relationship between the public and the private thresholds. The following theorem states this relationship.

Theorem 7. *In the observable setting,*
(i) For any $n_2 \geq 0$, $n_1(n_2) \geq \tilde{n}_1(n_2)$.
(ii) For any $n_1 \geq 0$, $\tilde{n}_2(n_1) \geq n_2(n_1)$.

Proof. First note that if (ii) holds, then (i) is also true. (ii) implies that for any state (n_1, n_2), whenever private is preferable in the decentralized setting, it is also preferable in the centralized setting. Equivalent to (i), which implies that for any state (n_1, n_2), whenever public is preferable in the centralized setting, it is also preferable in the decentralized setting.
To prove (ii), we have to prove the statement below:

$$c\frac{n_1 + 1}{\mu_1} > c\frac{n_2 + 1}{\mu_2} + f \Rightarrow \tilde{C}(n_1 + 1, n_2) > \tilde{C}(n_1, n_2 + 1) + f$$

It will be sufficient to show the following:

$$\tilde{C}(n_1+1,n_2)-\tilde{C}(n_1,n_2+1)\geq c\left(\frac{n_1+1}{\mu_1}-\frac{n_2+1}{\mu_2}\right)$$

Defining $\tilde{C}_0(n_1,n_2)$ as

$$\tilde{C}_0(n_1,n_2)=c\left(\frac{n_1+1}{\mu_1}-\frac{n_2+1}{\mu_2}\right)$$

We have

$$c\frac{n_1+1}{\mu_1}>c\frac{n_2+1}{\mu_2}+f\Rightarrow\tilde{C}_0(n_1,n_2)=c\left(\frac{n_1+1}{\mu_1}-\frac{n_2+1}{\mu_2}\right)>f$$

The claim holds for $\tilde{C}_0(n_1,n_2)$, and by Lemma 1, $\tilde{C}_t(n_1+1,n_2)-\tilde{C}_t(n_1,n_2+1)$ is increasing in t, so the proof is complete. □

4.6 Numerical Example

To clarify the model and show the thresholds, we analyse a simple numerical example. We assume that the buffer size for the public centre is 6 ($m_1=6$), and it is 4 for the private centre ($m_2=4$). The remaining parameters of the problem are $\lambda=4$, $\mu_1=2$, $\mu_2=3$, $c=0.5$, $f=1$, $c_B=2$.

The optimal policy for decentralized and centralized settings are shown in Figure 4.2 and Figure 4.3 respectively. According to these figures, we can see how the policy changes from decentralized setting to the centralized one. As we proved in Theorem 7, the public thresholds for centralized setting are smaller than the decentralized setting, while private thresholds are bigger. The comparison of public and private thresholds for decentralized and centralized settings are shown in Figure 4.4 and Figure 4.5.

We calculate the expected total cost rate in the system as below:

$$C=c\left(\sum_i i\pi_{i,j}+\sum_j j\pi_{i,j}\right)+f\lambda\pi_p+\lambda c_B\pi_{m_1,m_2}$$

Table 4.1: Observable Decentralized vs. Observable Centralized- Example with Parameters λ=4, μ_1=2, μ_2=3, c=0.5, f=1, c_B=2, m_1=6, m_2=4

	Decentralized	Centralized	% of Change
Probability of Joining Private	0.454	0.532	14.6%
Probability of Blocking	0.055	0.031	-77.4%
Public Expected Waiting Time	1.892	1.174	-61.2%
Private Expected Waiting Time	0.684	0.714	4.2%
Served Patients' Expected Cost Rate	4.503	3.988	-12.9%
Total Expected Cost Rate	4.942	4.234	-16.7%

while π_p is the probability that an individual joins the private centre.

Served patients' expected cost rate is the expected cost of individuals who joined the system, either in the public or private centre:

$$C_S = c\left(\sum_i i\pi_{i,j} + \sum_j j\pi_{i,j}\right) + f\lambda\pi_p$$

The expected waiting cost per person at the public centre is

$$W_1 = \frac{L_1}{\lambda p},$$

where p is the probability of joining the public centre, and L_1 is the expected number of individuals in the public centre per unit time.

The expected waiting cost per person at the private centre is

$$W_2 = \frac{L_2}{\lambda(1-p)},$$

where L_2 is the expected number of individuals in the private centre per unit time.

Table 4.1 shows some changes from decentralized setting to the centralized one in the observable case. We can see that moving from decentralized setting to the centralized one, the total expected cost rate, the cost rate of those who joined the system, and the blocking probability decrease, while the probability of joining the private centre increases. From Theorem 7, we know that for each n_1, $\tilde{n}_2(n_1)$ is greater than or equal

to $n_2(n_1)$. Equivalently, the thresholds for the private centre in the centralized setting are greater than or equal to the ones in the decentralized setting.

In the centralized setting the probability of joining the private centre is higher, but the central control reduces the cost by balancing the system. Furthermore, the expected waiting time per person in the public centre is lower in the centralized setting in comparison to the decentralized one, while for the private centre, it is higher.

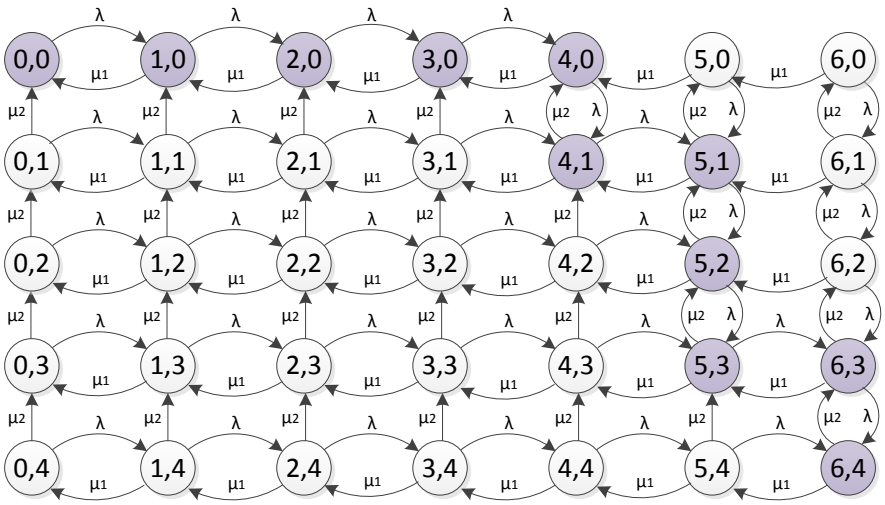

Figure 4.2: Optimal Policy in Observable Decentralized Setting- Example with Parameters λ=4, μ_1=2, μ_2=3, c=0.5, f=1, c_B=2, m_1=6, m_2=4

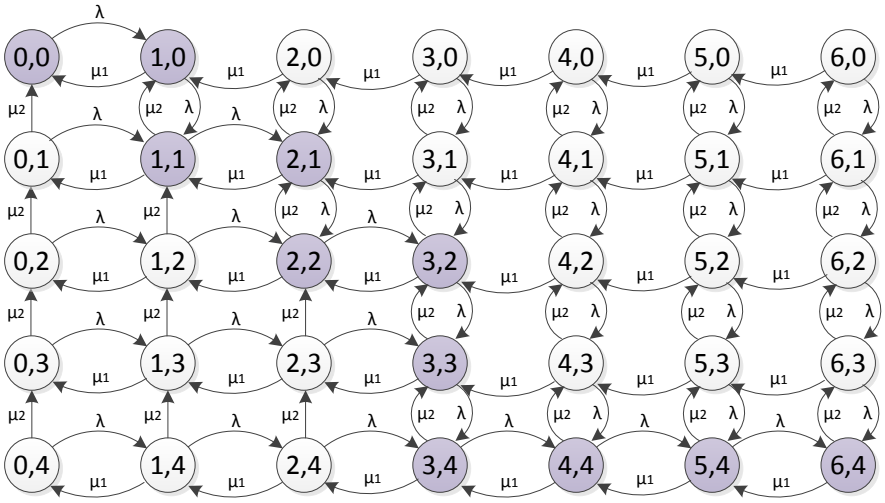

Figure 4.3: Optimal Policy in Observable Centralized Setting- Example with Parameters λ=4, μ_1=2, μ_2=3, c=0.5, f=1, c_B=2, m_1=6, m_2=4

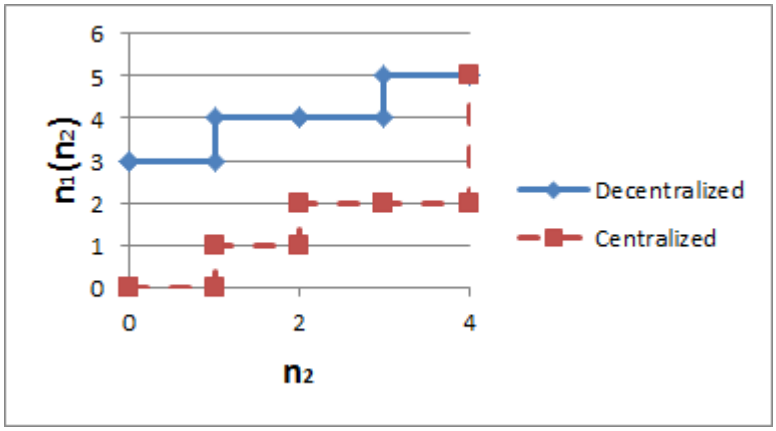

Figure 4.4: Comparison of Public Thresholds in Observable Decentralized and Centralized Settings

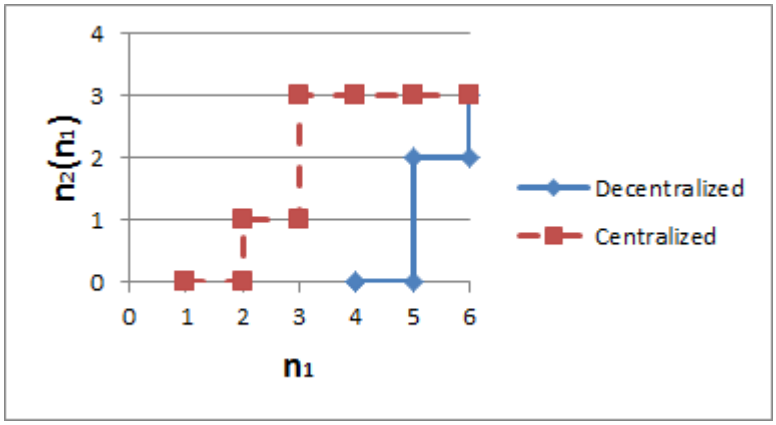

Figure 4.5: Comparison of Private Thresholds in Observable Decentralized and Centralized Settings

Chapter 5

UNOBSERVABLE CASE

This chapter assumes that the decision makers (whether individuals or the central authority) cannot see how many patients are present in each centre at the time of the decision. Practically, it is hard to make the queues observable in healthcare services, so many healthcare systems are unobservable. As before, we assume that balking is not allowed, individuals who join the system do not renege, and

(A1) $\mu_1 < \mu_2$,

(A2) $\lambda < \mu_1 + \mu_2$.

5.1 Decentralized Setting with Infinite Buffer

In the decentralized setting, patients individually decide which centre to attend without observing the number of patients in each centre. This can be modelled as a simultaneous-move game between patients. Let $\Gamma(\lambda, \mu_1, \mu_2, f)$ denote the game that corresponds to the unobservable decentralized model with parameters λ, μ_1, μ_2, f. Each patient chooses a probability $p \in [0, 1]$ to visit the public centre. Balking is not allowed, so the patient goes to the private centre with probability $1 - p$. The expected cost resulting from choosing p is the sum of the expected waiting cost and the fee charged by the centre chosen. This section restricts attention to symmetric strategy profiles.

Let $C(p, q)$ denote the expected cost of a patient who chooses to visit the public centre with probability $0 \leq p \leq 1$ while all other patients choose to visit it with probability $0 \leq q \leq 1$. In this case, the arrivals to the public and private centres form

two independent Poisson processes with rates $q\lambda$ and $(1-q)\lambda$, respectively.

$$C(p,q) = p\left(\frac{c}{(\mu_1 - q\lambda)^+}\right) + (1-p)\left(f + \frac{c}{(\mu_2 - (1-q)\lambda)^+}\right). \tag{5.1}$$

A strategy $p \in [0,1]$ is a best response to $q \in [0,1]$ if

$$C(p,q) \leq C(p',q) \quad \text{for all } p' \in [0,1]. \tag{5.2}$$

The equations (5.1) and (5.2) yield the following best-response point-to-set mapping

$$BR(q) \equiv \begin{cases} \{1\} & \text{if } \frac{c}{(\mu_1 - q\lambda)^+} < f + \frac{c}{(\mu_2-(1-q)\lambda)^+}, \\ \{0\} & \text{if } \frac{c}{(\mu_1 - q\lambda)^+} > f + \frac{c}{(\mu_2-(1-q)\lambda)^+}, \\ [0,1] & \text{if } \frac{c}{(\mu_1 - q\lambda)^+} = f + \frac{c}{(\mu_2-(1-q)\lambda)^+}. \end{cases} \tag{5.3}$$

A symmetric strategy profile that assigns p^* to every patient is a *(symmetric) Nash equilibrium* of $\Gamma(\lambda, \mu_1, \mu_2, f)$ if $p^* \in BR(p^*)$.

Theorem 8. *Under the assumptions (A1) and (A2), $\Gamma(\lambda, \mu_1, \mu_2, f)$ has a unique symmetric Nash equilibrium p_u^I. In particular,*

$$p_u^I = \begin{cases} 1 & \text{if } \frac{1}{(\mu_1-\lambda)^+} - \frac{1}{\mu_2} \leq \frac{f}{c}, \\ 0 & \text{if } \frac{1}{\mu_1} - \frac{1}{(\mu_2-\lambda)^+} > \frac{f}{c}, \\ p_u^I & \text{if } \frac{1}{\mu_1} - \frac{1}{(\mu_2-\lambda)^+} < \frac{f}{c} < \frac{1}{(\mu_1-\lambda)^+} - \frac{1}{\mu_2}, \end{cases} \tag{5.4}$$

where

$$p_u^I = \frac{-2c + f\left(\lambda + \mu_1 - \mu_2\right) + \sqrt{4c^2 + f^2(\mu_1 + \mu_2 - \lambda)^2}}{2f\lambda} \in (0,1), \tag{5.5}$$

and $p_u^I \in (0,1)$.

Proof. If $\frac{1}{(\mu_1-\lambda)^+} - \frac{1}{\mu_2} \leq \frac{f}{c}$, then (5.3) implies $1 \in BR(1)$, so $p_u^I = 1$ is a symmetric Nash equilibrium. If $\frac{1}{\mu_1} - \frac{1}{(\mu_2-\lambda)^+} \geq \frac{f}{c}$, then (5.3) implies $0 \in BR(0)$, so $p_u^I = 0$ is a

symmetric Nash equilibrium. In the remaining case,

$$\frac{1}{\mu_1} - \frac{1}{(\mu_2 - \lambda)^+} < \frac{f}{c} < \frac{1}{(\mu_1 - \lambda)^+} - \frac{1}{\mu_2} \tag{5.6}$$

$1 \notin BR(1)$ and $0 \notin BR(0)$ cannot be equilibrium strategies, so an equilibrium strategy (if there exists any) must satisfy

$$\frac{c}{(\mu_1 - p_u^I \lambda)^+} = f + \frac{c}{(\mu_2 - (1 - p_u^I)\lambda)^+}.$$

First note that for the equality to hold, either both denominators must be zero or both must be positive. The former implies that

$$\mu_1 \leq p\lambda \quad \text{and} \quad \mu_2 \leq (1-p)\lambda$$

Adding the two inequalities yields $\mu_1 + \mu_2 \leq \lambda$ which contradicts (A2). Hence if there exists a solution of (5.6), this solution should satisfy

$$\frac{c}{\mu_1 - p\lambda} = f + \frac{c}{\mu_2 - (1-p)\lambda},$$

$$\mu_1 \geq p\lambda,$$

$$\mu_2 \geq (1-p)\lambda.$$

Equivalently,

$$g(p) \equiv \frac{f\lambda^2}{c} p^2 + \lambda \left[\frac{f\left(\mu_2 - \mu_1 - \lambda\right)}{c} + 2 \right] p + \left[\mu_2 - \mu_1 - \lambda - \frac{f\mu_1\left(\mu_2 - \lambda\right)}{c} \right] = 0.$$

The solution is a Nash equilibrium if it is in the interval $(1 - \frac{\mu_2}{\lambda}, \frac{\mu_1}{\lambda}) \cap [0, 1]$.
Case 1: $\mu_1 < \lambda < \mu_2$. In this case, $(1 - \frac{\mu_2}{\lambda}, \frac{\mu_1}{\lambda}) \cap [0, 1] = [0, \frac{\mu_1}{\lambda}]$. Also the condition $\frac{1}{\mu_1} - \frac{1}{\mu_2 - \lambda} < \frac{f}{c}$ implies

$$\mu_2 - \lambda - \mu_1 - \frac{f}{c}\mu_1(\mu_2 - \lambda) < 0,$$

so the constant term of $g(p)$ is negative. Since the coefficient of the quadratic term is positive, $g(p)$ has a unique positive root, which is indeed p_u^I given by (5.5). Furthermore, since $g(0) < 0$ and

$$g\left(\frac{\mu_1}{\lambda}\right) = \mu_1 + \mu_2 - \lambda > 0$$

by (A2), the unique solution p_u^I is in the interval $(0, \frac{\mu_1}{\lambda})$ and is a Nash equilibrium.

Case 2: $\lambda < \mu_1 < \mu_2$. In this case, $(1 - \frac{\mu_2}{\lambda}, \frac{\mu_1}{\lambda}) \cap [0, 1] = [0, 1]$. Also the condition $\frac{f}{c} < \frac{1}{\mu_1 - \lambda} - \frac{1}{\mu_2}$ implies

$$\frac{f}{c}\mu_2(\mu_1 - \lambda) - \mu_2 + \mu_1 - \lambda < 0,$$

so

$$g(1) = \frac{f}{c}\mu_2(\lambda - \mu_1) + \lambda + \mu_2 - \mu_1 > 0.$$

Since $g(0) < 0$ and $g(1) > 0$, the unique solution p_u^I is in the interval $(0, 1)$ and so is a Nash equilibrium.

Case 3: $\mu_1 < \mu_2 \leq \lambda$. It is already shown that $g(\frac{\mu_1}{\lambda}) > 0$.

$$g\left(1 - \frac{\mu_2}{\lambda}\right) = \lambda - \mu_2 - \mu_1 < 0$$

by (A2), so $g(p)$ must have a unique root in $(1 - \frac{\mu_2}{\lambda}, \frac{\mu_1}{\lambda})$.

Note that $p_u^I \leq \frac{-2c + f(\lambda + \mu_1 - \mu_2) + 2c + f(\mu_1 + \mu_2 - \lambda)}{2f\lambda} = \frac{\mu_1}{\lambda}$ is the largest root of $g(p)$, so $p_I \in (1 - \frac{\mu_2}{\lambda}, \frac{\mu_1}{\lambda})$ is a Nash equilibrium.

This completes the proof of existence of a symmetric Nash equilibrium for all possible cases. To show uniqueness for the rest of the cases, note that if $\frac{f}{c} \geq \frac{1}{(\mu_1 - \lambda)^+} - \frac{1}{\mu_2}$, then $\mu_1 > \lambda$, so $\mu_2 > \lambda$ and $0 \notin BR(0)$. In this case, $\frac{f}{c} > \frac{1}{\mu_1} - \frac{1}{\mu_2 - \lambda}$, so $g(p)$ has a negative constant term and $g(1) < 0$, so the unique positive root of $g(p)$ is greater than 1. If $\frac{f}{c} < \frac{1}{\mu_1} - \frac{1}{(\mu_2 - \lambda)^+}$, then $\mu_2 > \lambda$, the constant term and the coefficient of the linear term in $g(p)$ are both positive, so $g(p)$ does not have any positive roots. □

The following Corollary states the monotonicity of p_u^I in different parameters of the model.

Corollary 6. *The equilibrium probability of joining the public centre, p_u^I is increasing in f and μ_1, and decreasing in c and μ_2.*

Proof. Let $z = \frac{c}{f}$ and write (5.5) as

$$h(z, \mu_1, \mu_2) = -z + \frac{1}{2}(\lambda + \mu_1 - \mu_2) + \sqrt{z^2 + \frac{1}{4}(\mu_1 + \mu_2 - \lambda)^2}.$$

Then

$$\frac{dh(z, \mu_1, \mu_2)}{dz} = -1 + \frac{z}{\sqrt{z^2 + k^2}} < 0,$$

where $k = \frac{1}{2}(\mu_1 + \mu_2 - \lambda)$, so p_u^I is decreasing in $\frac{c}{f}$, increasing in f and decreasing in c. To show monotonicity in μ_1 and μ_2,

$$\begin{aligned}\frac{dh(z, \mu_1, \mu_2)}{d\mu_2} = -\frac{1}{2} + \frac{\frac{1}{4}(\mu_1 + \mu_2 - \lambda)}{\sqrt{z^2 + \frac{1}{4}(\mu_1 + \mu_2 - \lambda)^2}} &= -\frac{1}{2}\left(1 - \frac{\frac{1}{2}(\mu_1 + \mu_2 - \lambda)}{\sqrt{z^2 + \frac{1}{4}(\mu_1 + \mu_2 - \lambda)^2}}\right) \\ &= -\frac{1}{2}\left(1 - \frac{k}{\sqrt{z^2 + k^2}}\right) < 0,\end{aligned}$$

$$\frac{dh(z, \mu_1, \mu_2)}{d\mu_1} = \frac{1}{2} + \frac{\frac{1}{4}(\mu_1 + \mu_2 - \lambda)}{\sqrt{z^2 + \frac{1}{4}(\mu_1 + \mu_2 - \lambda)^2}} > 0,$$

where the last inequality follows from (A2). □

The following example illustrates nonmonotonicity of p_u^I in λ.

Example: Let $\mu_1 = 2$, $\mu_2 = 3$, $f = 1$, $c = 0.5$. Then $p_u^I \simeq 0.427$ for $\lambda = 4$, $p_u^I \simeq 0.4$ for $\lambda = 5$, and $p_u^I \simeq 0.451$ for $\lambda = 6$.

The monotonicity results stated in Corollary 6 are intuitive. As the rate of service in the public centre, or the fee of joining the private centre increases, a higher proportion of patients tend to join the public centre. On the other hand, as the rate of service in the private centre, or the waiting cost per unit time per person increases, the probability that patients join the public centre decreases. A higher waiting cost per unit time implies a higher displeasure for waiting, so more individuals tend to pay the fee and join the private centre instead of waiting more in the public one.

5.2 Decentralized Setting with Finite Buffer

In reality, several healthcare systems have a limited buffer. When the number of patients in the system reaches to a specific number, new arrivals are not admitted to the system. These patients have to go to other healthcare centres, or take the risk of not being treated. In both cases, the cost of being blocked is considerable. Suppose that the buffer size of the public centre is m_1, and it is m_2 for the private centre. The blocking cost is assumed to be c_B. As reneging is not allowed, once a patient is blocked at one centre, she cannot join the other centre. Similar to the case with infinite buffer, this situation can be modelled as a simultaneous-move game between patients, and each patient acts in order to minimize her expected cost.

To calculate the expected cost of an individual, and also the total expected cost rate of the system, we use the formulas for M/M/1/c queues [38]. For a queue with buffer size N, the expected number of individuals in the system per unit of time is

$$L = \frac{\rho}{1-\rho} - \frac{(N+1)\rho^{N+1}}{1-\rho^{N+1}}, \tag{5.7}$$

where $\rho = \frac{\lambda}{\mu}$. The probability that an arrival is blocked is

$$p_N = \frac{\rho^N(1-\rho)}{1-\rho^{N+1}}, \tag{5.8}$$

and the effective arrival rate, which is the rate of arrivals of those who could enter the system is

$$\bar{\lambda} = \lambda(1-p_N). \tag{5.9}$$

By Little's Law,

$$W = \frac{L}{\bar{\lambda}}$$

If individuals join the public centre with probability p, then the arrivals to the public and private centres form two independent Poisson processes with rates λp and $\lambda(1-p)$, respectively.

Using equation (5.7), the expected number of individuals per unit time in the public and private centres are

$$L_1(p) = \frac{\frac{\lambda p}{\mu_1}}{1-\frac{\lambda p}{\mu_1}} - \frac{(m_1+1)\left(\frac{\lambda p}{\mu_1}\right)^{m_1+1}}{1-\left(\frac{\lambda p}{\mu_1}\right)^{m_1+1}}, \tag{5.10}$$

$$L_2(p) = \frac{\frac{\lambda(1-p)}{\mu_2}}{1-\frac{\lambda(1-p)}{\mu_2}} - \frac{(m_2+1)\left(\frac{\lambda(1-p)}{\mu_2}\right)^{m_2+1}}{1-\left(\frac{\lambda(1-p)}{\mu_2}\right)^{m_2+1}}, \tag{5.11}$$

respectively. Using Little's law, the expected waiting time per person in the public and private centres are

$$W_1(p) = \frac{L_1(p)}{\lambda p(1-p_{m_1})} = \frac{\frac{\lambda p}{\mu_1}\left(1-\left(\frac{\lambda p}{\mu_1}\right)^{m_1+1}\right)}{\lambda p\left(1-\frac{\lambda p}{\mu_1}\right)\left(1-\left(\frac{\lambda p}{\mu_1}\right)^{m_1}\right)} - \frac{(m_1+1)\left(\frac{\lambda p}{\mu_1}\right)^{m_1+1}}{\lambda p\left(1-\left(\frac{\lambda p}{\mu_1}\right)^{m_1}\right)}, \tag{5.12}$$

$$\begin{aligned} W_2(p) &= \frac{L_2(p)}{\lambda(1-p)(1-p_{m_2})} \qquad (5.13)\\ &= \frac{\frac{\lambda(1-p)}{\mu_2}\left(1-\left(\frac{\lambda(1-p)}{\mu_2}\right)^{m_2+1}\right)}{\lambda(1-p)\left(1-\frac{\lambda(1-p)}{\mu_2}\right)\left(1-\left(\frac{\lambda(1-p)}{\mu_2}\right)^{m_2}\right)} - \frac{(m_2+1)\left(\frac{\lambda(1-p)}{\mu_2}\right)^{m_2+1}}{\lambda(1-p)\left(1-\left(\frac{\lambda(1-p)}{\mu_2}\right)^{m_2}\right)}, \end{aligned}$$

respectively.

Let $C_1(q)$ and $C_2(q)$ be the expected cost of joining the public and the private centres, respectively. Then

$$C_1(q) = [1-p_{B_1}(q)]\, cW_1(q) + p_{B_1}(q)c_B,$$

$$C_2(q) = [1-p_{B_2}(q)]\, [cW_2(q)+f] + p_{B_2}(q)c_{B_2}.$$

Letting p_{B_1} and p_{B_2} be the blocking probabilities for the public and the private centres, respectively, we have

$$C(p,q) = p\left[(1-p_{B_1}(q)cW_1(q) + p_{B_1}(q)c_B\right] + (1-p)\left[(1-p_{B_2}(q))(cW_2(q)+f) + p_{B_2}(q)c_{B_2}\right] \tag{5.14}$$

A strategy $p \in [0,1]$ is a best response to $q \in [0,1]$ if

$$C(p,q) \leq C(p',q) \quad \text{for all } p' \in [0,1]. \tag{5.15}$$

The best-response point-to-set mapping is then

$$BR(q) \equiv \begin{cases} \{1\} & \text{if } C_1(1) < C_2(1), \\ \{0\} & \text{if } C_1(0) > C_2(0), \\ [0,1] & \text{if } C_1(q) = C_2(q), \end{cases} \tag{5.16}$$

which is equivalent to

$$BR(q) = \begin{cases} \{1\} & \text{if } \frac{L_1(1)}{\lambda} + p_{B_1}(1)\frac{c_B}{c} < \frac{1}{\mu_2} + \frac{f}{c}, \\ \{0\} & \text{if } \frac{L_2(0)}{\lambda} + p_B(0)\frac{c_{B_2}}{c} < \frac{1}{\mu_1} - \frac{f}{c}\left(1-p_{B_2}(0)\right), \\ [0,1] & \text{if } c\frac{L_1(p)}{\lambda p} + p_{B_1}(p)c_B = c\frac{L_2(p)}{\lambda(1-p)} + f\left(1-p_{B_2}(p)\right) + p_{B_2}(p)c_B. \end{cases} \tag{5.17}$$

Theorem 9. *Under the assumptions (A1) and (A2), $\Gamma(\lambda, \mu_1, \mu_2, f)$ has a symmetric Nash equilibrium p_u^I. In particular,*

$$p_u^I = \begin{cases} 1 & \text{if } \frac{L_1(1)}{\lambda} + p_{B_1}(1)\frac{c_B}{c} - \frac{1}{\mu_2} \leq \frac{f}{c}, \\ 0 & \text{if } \frac{1}{(1-p_{B_2}(0))}\left(\frac{1}{\mu_1} - \frac{L_2(0)}{\lambda} - p_B(0)\frac{c_{B_2}}{c}\right) > \frac{f}{c}, \\ p_u^I & \text{if } \frac{1}{(1-p_{B_2}(0))}\left(\frac{1}{\mu_1} - \frac{L_2(0)}{\lambda} - p_B(0)\frac{c_{B_2}}{c}\right) < \frac{f}{c} < \frac{L_1(1)}{\lambda} + p_{B_1}(1)\frac{c_B}{c} - \frac{1}{\mu_2}, \end{cases} \tag{5.18}$$

where p_u^I *is the root of the equation below*

$$c\left(\frac{1}{\mu_1-\lambda p}-\frac{(m_1+1)\left(\frac{\lambda p}{\mu_1}\right)^{m_1+1}}{\lambda p\left(1-\left(\frac{\lambda p}{\mu_1}\right)^{m_1+1}\right)}\right)+c_B\frac{\left(\frac{\lambda p}{\mu_1}\right)^{m_1}\left(1-\frac{\lambda p}{\mu_1}\right)}{1-\left(\frac{\lambda p}{\mu_1}\right)^{m_1+1}}=$$

$$c\left(\frac{1}{\mu_2-\lambda(1-p)}-\frac{(m_2+1)\left(\frac{\lambda(1-p)}{\mu_2}\right)^{m_2+1}}{\lambda(1-p)\left(1-\left(\frac{\lambda(1-p)}{\mu_2}\right)^{m_2+1}\right)}\right)+$$

$$c_B\frac{\left(\frac{\lambda(1-p)}{\mu_2}\right)^{m_2}\left(1-\frac{\lambda(1-p)}{\mu_2}\right)}{1-\left(\frac{\lambda(1-p)}{\mu_2}\right)^{m_2+1}}+f\left(\frac{1-\left(\frac{\lambda(1-p)}{\mu_2}\right)^{m_2}}{1-\left(\frac{\lambda(1-p)}{\mu_2}\right)^{m_2+1}}\right)$$

and $p_u^I\in(0,1)$.

Proof. If $\frac{L_1(1)}{\lambda}+p_{B_1}(1)\frac{c_B}{c}-\frac{1}{\mu_2}\leq\frac{f}{c}$, then (5.17) implies $1\in BR(1)$, so $p_u^I=1$ is a symmetric Nash equilibrium. If $\frac{1}{(1-p_{B_2}(0))}\left(\frac{1}{\mu_1}-\frac{L_2(0)}{\lambda}-p_B(0)\frac{c_{B_2}}{c}\right)>\frac{f}{c}$, then (5.17) implies $0\in BR(0)$, so $p_u^I=0$ is a symmetric Nash equilibrium. In the remaining case,

$$\frac{1}{(1-p_{B_2}(0))}\left(\frac{1}{\mu_1}-\frac{L_2(0)}{\lambda}-p_B(0)\frac{c_{B_2}}{c}\right)<\frac{f}{c}<\frac{L_1(1)}{\lambda}+p_{B_1}(1)\frac{c_B}{c}-\frac{1}{\mu_2},$$

$1\notin BR(1)$ and $0\notin BR(0)$ cannot be equilibrium strategies, so an equilibrium strategy (if there exists any) must satisfy

$$c\frac{L_1(p)}{\lambda p}+p_{B_1}(p)c_B=c\frac{L_2(p)}{\lambda(1-p)}+f(1-p_{B_2}(p))+p_{B_2}(p)c_B. \tag{5.19}$$

If equation (5.19) has a unique root in the interval of $(0,1)$, the proof is complete. The existence and uniqueness of this root is not proved, although we did not see any counterexample. □

5.3 Centralized Setting with Infinite Buffer

This section assumes the presence of a central authority that sets a policy to be adopted by every individual in the system. In the unobservable case, the authority does not observe the state of each centre prior to assigning individuals to the public and private centres. The objective of the authority is to minimize the total expected cost of all patients visiting the system. Let $\tilde{C}(p)$ denote the total expected cost if every incoming patient chooses the public centre with probability $0 \leq p \leq 1$ and the private one with probability $1-p$. Since the interarrival and service times are exponentially distributed,

$$\tilde{C}(p) = \lambda \left[p \left(\frac{c}{(\mu_1 - p\lambda)^+} \right) + (1-p) \left(f + \frac{c}{(\mu_2 - (1-p)\lambda)^+} \right) \right]. \tag{5.20}$$

Theorem 10. *Under (A1) and (A2), the unique minimizer of* $\tilde{C}(p)$ *on the interval* $[0,1]$ *is*

$$p_u^S = \begin{cases} 1 & \text{if } \frac{\mu_1}{((\mu_1-\lambda)^+)^2} - \frac{1}{\mu_2} \leq \frac{f}{c}, \\ 0 & \text{if } \frac{1}{\mu_1} - \frac{\mu_2}{((\mu_2-\lambda)^+)^2} > \frac{f}{c}, \\ p_u^S & \text{if } \frac{1}{\mu_1} - \frac{\mu_2}{((\mu_2-\lambda)^+)^2} < \frac{f}{c} < \frac{\mu_1}{((\mu_1-\lambda)^+)^2} - \frac{1}{\mu_2}, \end{cases} \tag{5.21}$$

where p_u^S *is the unique nonnegative root of the fourth-degree polynomial*

$$h(p) \equiv \frac{f}{c}(\mu_1 - p\lambda)^2(\mu_2 - (1-p)\lambda)^2 - \mu_1(\mu_2 - (1-p)\lambda)^2 + \mu_2(\mu_1 - p\lambda)^2. \tag{5.22}$$

and $p_u^S \in (0,1)$.

Proof. First consider the case where $\mu_1 > \lambda$. Differentiating (5.20) with respect to p

gives:

$$\frac{d\tilde{C}(p)}{dp} = c\left[\frac{\mu_1}{(\mu_1 - p\lambda)^2} - \frac{\mu_2}{(\mu_2 - (1-p)\lambda)^2}\right] - f, \tag{5.23}$$

$$\frac{d^2\tilde{C}(p)}{dp^2} = 2c\lambda\left[\frac{\mu_1}{(\mu_1 - p\lambda)^3} + \frac{\mu_2}{(\mu_2 - (1-p)\lambda)^3}\right]. \tag{5.24}$$

Under assumption (A1), (5.24) is strictly positive for $p \in [0,1]$, so (5.20) is strictly convex and has a unique minimizer in the interval $[0,1]$. The unique minimizer is $p_u^S = 0$ if (5.23) is nonnegative at $p = 0$, $p_u^S = 1$ if (5.23) is nonpositive at $p = 1$. Otherwise, the minimizer is in the interior $(0,1)$, found by equating (5.23) to zero, which yields the equation (5.22).

If $\mu_1 \leq \lambda < \mu_2$, then $\tilde{C}(p)$ is strictly convex in $p \in \left[0, \frac{\lambda}{\mu_1}\right]$ and equals infinity on $p \in \left[\frac{\lambda}{\mu_1}, 1\right]$, so its unique minimizer is

$$p_u^S \;=\; \begin{cases} 0 & \text{if } \frac{1}{\mu_1} - \frac{\mu_2}{(\mu_2-\lambda)^2} \geq \frac{f}{c}, \\ \\ p_u^S \in \left(0, \frac{\mu_1}{\lambda}\right) & \text{if } \frac{f}{c} > \frac{1}{\mu_1} - \frac{\mu_2}{(\mu_2-\lambda)^2}, \end{cases} \tag{5.25}$$

where p_u^S is the unique nonnegative root of the fourth-degree polynomial (5.22).

If $\mu_2 \leq \lambda$, then $\tilde{C}(p)$ is strictly convex in $p \in \left[1 - \frac{\mu_2}{\lambda}, \frac{\mu_1}{\lambda}\right]$ and equals infinity in the rest of the interval $[0,1]$, so its unique minimizer is

$$p_u^S = p_S \in \left(1 - \frac{\mu_2}{\lambda}, \frac{\mu_1}{\lambda}\right) \tag{5.26}$$

where p_u^S is the unique nonnegative root of the fourth-degree polynomial (5.22). □

Example: Free Private Centre. Letting $f = 0$ in (5.22) gives

$$h(p) = \mu_2(\mu_1 - p\lambda)^2 - \mu_1(\mu_2 - (1-p)\lambda)^2.$$

For $\mu_1 > \lambda$, $h(p) = 0$ has the unique solution:

$$0 \leq p_u^S = \frac{\lambda - \sqrt{\mu_2}\left(\sqrt{\mu_2} - \sqrt{\mu_1}\right)}{\lambda\left(\sqrt{\frac{\mu_2}{\mu_1}} + 1\right)} \leq \frac{1}{2}. \tag{5.27}$$

Hence when the private centre provides free service at a faster rate than the public one, the central authority assigns the majority of the patients to the private centre. In particular, it assigns all the patients to the private centre, i.e., $p_u^S = 0$ if and only if $\lambda = \mu_2 - \sqrt{\mu_1\mu_2}$. Let $\alpha \equiv \sqrt{\frac{\mu_2}{\mu_1}} > 1$, then

$$p_u^S = \frac{\lambda - \alpha^2\mu_1 + \alpha\mu_1}{\lambda\left(\alpha + 1\right)},$$

which is decreasing in $\alpha > 1$. This implies that as the gap between the service rates of the public and the private centre increases, the proportion of patients visiting the private centre increases.

5.4 Centralized Setting with Finite Buffer

Suppose that the buffer size for the public centre is m_1, and it is m_2 for the private centre. The blocking cost is c_B. By (5.8) and (5.9), The effective arrival rate for the public centre is

$$\lambda_1(p) = \lambda p\left(1 - \frac{(\frac{p\lambda}{\mu_1})^{m_1}(1 - \frac{\lambda p}{\mu_1})}{1 - (\frac{\lambda p}{\mu_1})^{m_1+1}}\right), \tag{5.28}$$

The effective arrival rate for the private centre is

$$\lambda_2(p) = (1-p)\lambda\left(1 - \frac{(\frac{\lambda(1-p)}{\mu_2})^{m_2}(1 - \frac{\lambda(1-p)}{\mu_2})}{1 - (\frac{\lambda(1-p)}{\mu_2})^{m_2+1}}\right), \tag{5.29}$$

The total expected cost rate is

$$TC(p) = c(L_1(p) + L_2(p)) + f\lambda_2(p) + \lambda c_B p_B(p), \tag{5.30}$$

and the blocking probability for the system is

$$p_B = pp_{B_1} + (1-p)p_{B_2}, \tag{5.31}$$

where p_{B_1} and p_{B_2} are the probabilities that the public centre and the private centre are blocked, respectively. These probabilities can be calculated by (5.8).
In this case, the total expected cost function is not always convex. If it is convex, by Theorem 10, we have a unique minimizer of $TC(p)$ on the interval $[0, 1]$.

5.5 Centralized vs. Decentralized Setting

Example: Similar to the previous case, suppose that $m_1 = 6$, $m_2 = 4$, and the remaining parameters are $\lambda = 4$, $\mu_1 = 2$, $\mu_2 = 3$, $c = 0.5$, $f = 1$, and $c_B = 2$.
Table 5.1 shows some changes from decentralized setting to the centralized one in the unobservable setting. As it was mentioned, the total expected cost function is not necessarily convex. For this numerical example, it is convex as it is shown in Figure 5.1, so we can use our results to find the optimal answers.

Table 5.1: Unobservable Decentralized vs. Unobservable Centralized- Example with Parameters λ=4, μ_1=2, μ_2=3, c=0.5, f=1, c_B=2, m_1=6, m_2=4

	Decentralized	Centralized	% of change
Probability of Joining Private	0.205	0.591	65.3%
Probability of Blocking	0.308	0.099	-211.1%
Public Expected Waiting Time	1.443	1.352	-5.8%
Private Expected Waiting Time	0.449	0.649	44.5%
Served Patients' Expected Cost Rate	3.294	3.965	16.9%
Total Expected Cost Rate	5.758	4.756	-21%

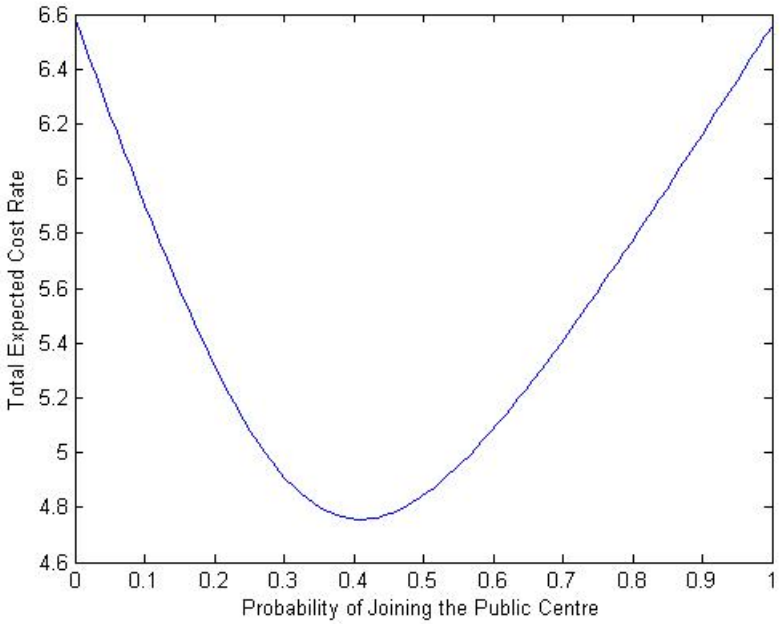

Figure 5.1: Total Expected Cost in Unobservable Setting- Example with Parameters λ=4, μ_1=2, μ_2=3, c=0.5, f=1, c_B=2, m_1=6, m_2=4

According to Table 5.1, the centralized setting sends a higher proportion of patients to the private centre to make the whole system more balanced. In the centralized case, the probability of blocking and the total expected cost rate are also lower, as in this case the objective is to minimize the total expected cost.

It can also be seen that in the centralized setting, the expected waiting time for each individual who joins the public centre is less than the decentralized setting, while

the expected waiting time per person in the private centre is higher. This happens because the centralized setting utilize the private centre more in order to decrease the probability of blocking. Sending more patients to the private centre also leads to a higher expected cost rate of the patients who enter the system and obtain the service.

Chapter 6

APPLICATION

As diagnostic imaging facilities are considered as one of the most wide-spreading and expensive healthcare technologies, it becomes increasingly important to manage the use of these resources in a more efficient way. In this chapter we consider a public and a private centre, each equipped with one MRI machine.

Green et al. [29] examine the management of MRI facility. They consider several patient types: inpatients, outpatients, and emergency patients. Each type has its own revenue, waiting cost, probability of arriving, and the penalty for not being served during one day. They design the outpatient appointment schedule, and establish a policy for admitting patients into the service.

We use some data from [29], and estimate other parameters. We then perform a sensitivity analysis to understand the effect of parameters on the whole system.

Green et al. assume that there are N identical service slots for each MRI machine per day. In each slot the arrival of each type of patients has a Bernoulli distribution with a specific probability according to the type of the patient. In our case, we assume that all patients are "outpatient".

In the base case, we consider the parameters as below:

$\lambda = 29$, $\mu_1 = 15$, $\mu_2 = 20$, $c = 120$, $c_B = 5000$, $f = 300$, $m_1 = m_2 = 50$.

According to [29], we consider 20 slots of 45 minutes for each MRI machine in the private centre, and 15 slots of the same length for public centre. The arrival rate is calculated as

$$(20 + 15)(0.84) = 29.4 \simeq 29$$

where 0.84 is the outpatient arrival probability from [29], which means in each slot an outpatient arrival could happen by a Bernoulli distribution with probability 0.84. So the arrival of the day is the sum of 35 Bernoulli slots which is a Binomial distribution with the mean of 29.4. We round the mean to 29.

The waiting cost per day is calculated based on [29] as

$$c = \frac{30000}{250} = 120$$

where 30000 is the average annual salary in dollars, and 250 is the number of working days in a year.

The expected cost of an MRI test is \$1000, which we assume on average 70% of this amount will be paid by the insurance, so each individual has to pay the fee of \$300 to enter the private centre ($f = 300$).

The blocking cost is assumed to be \$5000 which is higher than the corresponding holding cost if the patient was not blocked.

Finally, we assume that the buffer size of each centre is 50. We choose these buffer sizes so that the probability of blocking is neglectable for the system in different cases.

The Base Case

We consider three performance measures: The probability of joining the private centre, the probability of blocking, and the total expected cost rate. The performance measures for the base case are shown in Table 6.1. For this case, the expected waiting cost rate is also calculated, so we can see which amount of the total expected cost rate is caused by waiting in the system.

Table 6.1: Base Case

	Obs.Decent	Obs.Cent	Unobs.Decent	Unobs.Cent
Prob of joining private	0.483	0.498	0.459	0.552
Prob of Blocking	0.000	0.000	0.026	0.000
Expected Waiting Cost Rate	4882.739	835.924	4308.069	1252.327
Total Expected Cost Rate	9085	5170.98	12097.887	6062.733

The total expected cost rate function for the unobservable case is shown in Figure 6.1. For all cases of this chapter, this function is convex, so we can find the optimal solution in these cases by using our results.

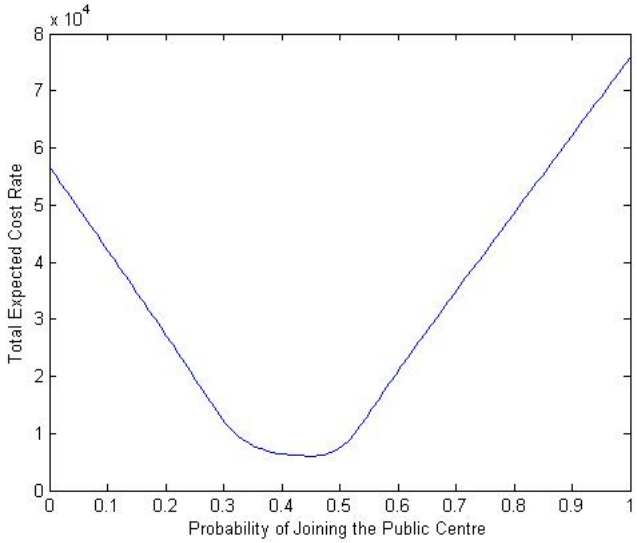

Figure 6.1: Total Expected Cost in Unobservable Setting- Base Case

According to the numbers in Table 6.1, as it was expected, the observable centralized case has the lowest total expected cost rate. It is shown that in all numerical examples of this chapter, the optimal case is the observable centralized case, which has the lowest total expected cost rate. The unobservable centralized setting, and observable decentralized setting have the second and the third lowest costs, respectively. The worst case is the unobservable decentralized one.

In the observable decentralized, and unobservable decentralized settings, the proportion of patients who join the private centre is less than the optimal proportion, while in the unobservable centralized case, the central authority sends more individuals to the private centre in comparison to the optimal case.

The probability of blocking is almost 0 for all cases, except for the unobservable decentralized setting. Another observation is that the ratio of the total expected waiting cost rate and the total expected cost rate is lower in the centralized cases. As

the probability of blocking is neglectable in these cases, the most proportion of the total expected cost rate is caused by paying the fee to join the private centre.

To perform the sensitivity analysis on the parameters λ, μ_1, f, c, and m_2, we decrease and increase the base case values by approximately 15%. We also perform the sensitivity analysis on c_B and increase it to 15% and 30% of its base value.

Sensitivity Analysis with respect to λ

We perform the sensitivity analysis for two values of λ, 25 and 34.

Table 6.2: Performance Measure for $\lambda = 25$

	Obs.Decent	Obs.Cent	Unobs.Decent	Unobs.Cent
Prob of joining private	0.4	0.424	0.377	0.492
Prob of Blocking	0.000	0.000	0.027	0.000
Total Expected Cost Rate	7540.78	3785.05	10177.695	4545.388

Table 6.3: Performance Measure for $\lambda = 34$

	Obs.Decent	Obs.Cent	Unobs.Decent	Unobs.Cent
Prob of joining private	0.553	0.565	0.532	0.573
Prob of Blocking	0.005	0.002	0.029	0.009
Total Expected Cost Rate	13581.5	9564.43	15755.047	11796.237

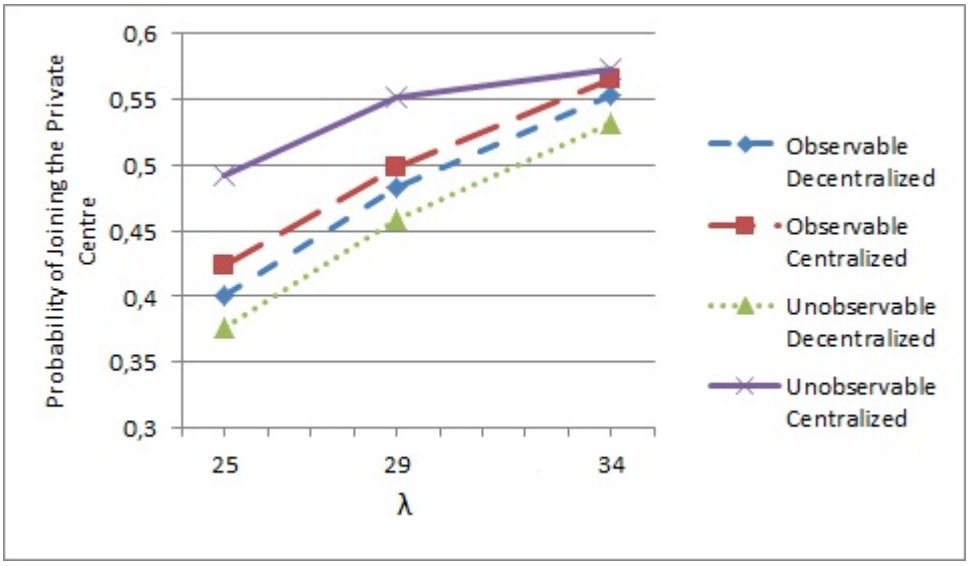

Figure 6.2: Sensitivity Analysis on Probability of Joining Private Centre with respect to λ

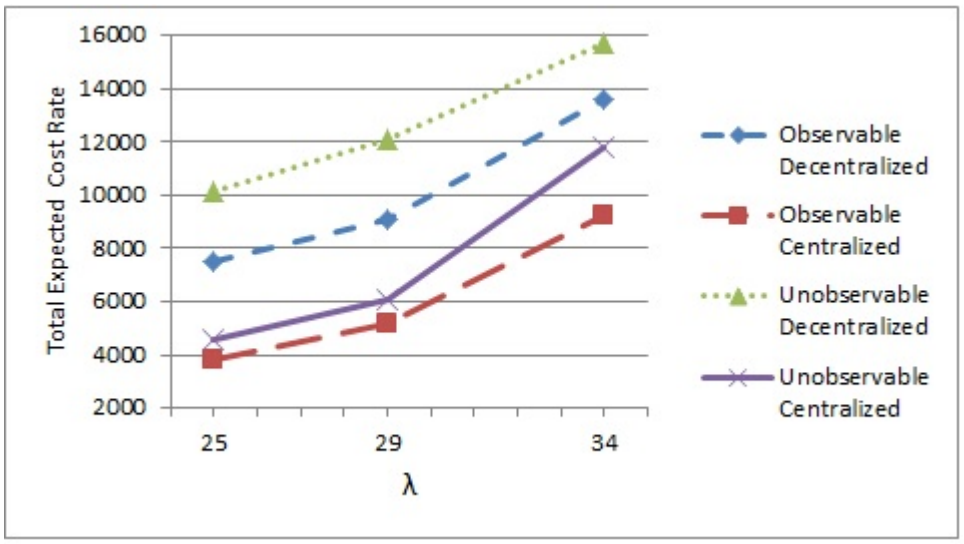

Figure 6.3: Sensitivity Analysis on Total Expected Cost with respect to λ

As λ increases, the probability of joining the private centre, the probability of blocking, and the total expected cost rate increase. According to Figure 6.2, for the unobservable centralized setting, the rate of increasing in the probability of joining the private centre with respect to λ is higher for lower values of λ. In this case, as

the system becomes more congested, increasing the arrival rate has a smaller effect on increasing the probability of sending a patient to the private centre. For the other three cases, the rate is almost constant. In Figure 6.2, we can also see that in the unobservable centralized case, the probability of sending individuals to the private centre is considerably higher than the remaining cases, and as λ increases, the differences of the four cases become smaller. According to Figure 6.3, for higher values of λ, the rate of changes in the total expected cost rate, with respect to λ is higher. In other words, for an already congested system, increasing the arrival rate leads to a higher increase in the total expected cost rate, in comparison to the cases with lower arrival rates. In Figure 6.3, we can also observe that for lower arrival rates, the difference between the total cost rate of the observable centralized and decentralized cases are smaller than the other cases, but as λ increases, this difference also increases. As λ increases, the system's load which is calculated as

$$\rho = \frac{\lambda}{\mu_1 + \mu_2}$$

changes from $\frac{25}{34} = 0.714$ to $\frac{29}{34} = 0.829$, and then to $\frac{34}{35} = 0.971$.

Sensitivity Analysis with respect to μ_1

We perform the sensitivity analysis for two values of μ_1, 13 and 17.

Table 6.4: Performance Measures for $\mu_1 = 13$

	Obs.Decent	Obs.Cent	Unobs.Decent	Unobs.Cent
Prob of joining private	0.552	0.566	0.529	0.603
Prob of Blocking	0.000	0.000	0.024	0.000
Total Expected Cost Rate	9390.1	5987.65	12679.13	7018.875

Table 6.5: Performance Measures for $\mu_1 = 17$

	Obs.Decent	Obs.Cent	Unobs.Decent	Unobs.Cent
Prob of joining private	0.414	0.433	0.389	0.495
Prob of Blocking	0.000	0.000	0.028	0.000
Total Expected Cost Rate	8893.66	4503.12	11605.67	5360.716

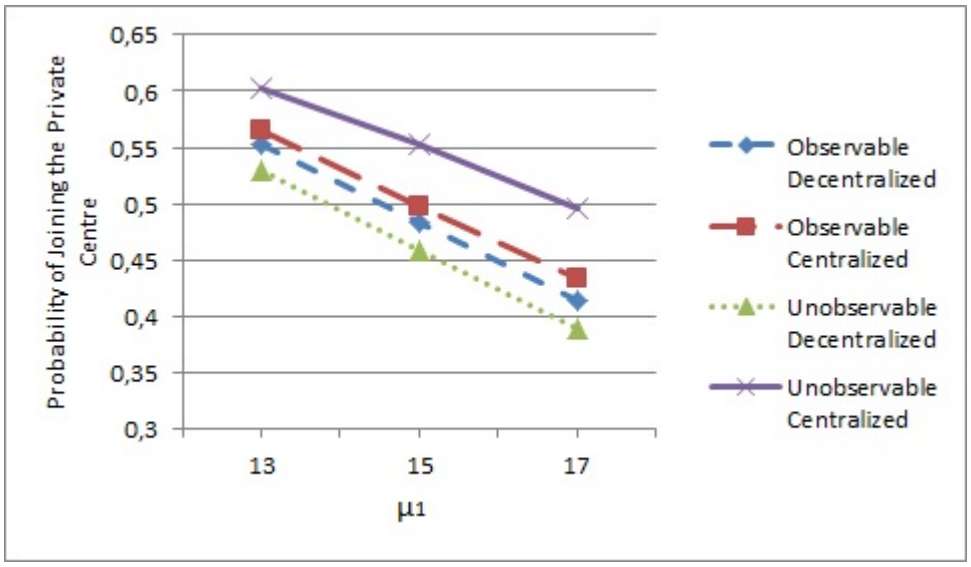

Figure 6.4: Sensitivity Analysis on Probability of Joining Private Centre with respect to μ_1

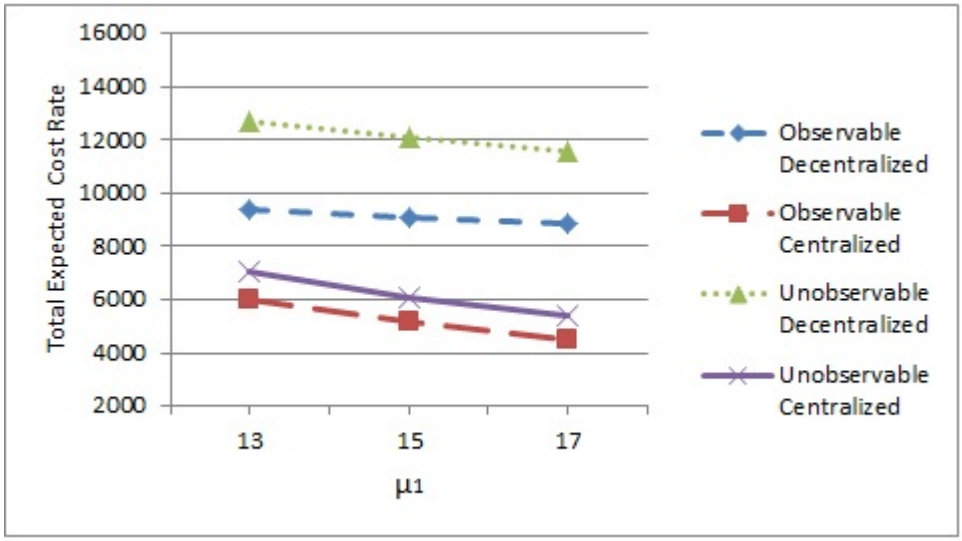

Figure 6.5: Sensitivity Analysis on Total Expected Cost with respect to μ_1

As μ_1 increases, the probability of joining the private centre increases, while the probability of blocking, and the total expected cost rate decrease. According to Figure 6.10, for the probability of joining the private centre, the rate of decreasing with respect to μ_1 is almost constant, as the system's load does not change significantly. As μ_1 increases, the load of the system changes from $\frac{29}{13+20} = 0.879$ to $\frac{29}{15+20} = 0.829$, and then to $\frac{29}{17+20} = 0.784$. According to Figure 6.11, for the total expected cost rate, the rate of decreasing with respect to μ_1 is almost the same for all cases, except for the observable centralized setting. For the observable centralized setting, which is the optimal setting, increasing the public service rate has a small effect on reducing the total expected cost of the system.

Sensitivity Analysis with respect to f

We perform the sensitivity analysis for two values of the insurance coverage, 60% and 80%. So different values of f would be 200 and 400.

Table 6.6: Performance Measures for $f = 200$

	Obs.Decent	Obs.Cent	Unobs.Decent	Unobs.Cent
Prob of joining private	0.483	0.51	0.472	0.559
Prob of Blocking	0.000	0.000	0.016	0.000
Total Expected Cost Rate	6259.47	3717.14	8886.775	4451.418

Table 6.7: Performance Measures for $f = 400$

	Obs.Decent	Obs.Cent	Unobs.Decent	Unobs.Cent
Prob of joining private	0.483	0.497	0.447	0.546
Prob of Blocking	0.000	0.000	0.037	0.000
Total Expected Cost Rate	11962.5	6613.9	15195.983	7655.166

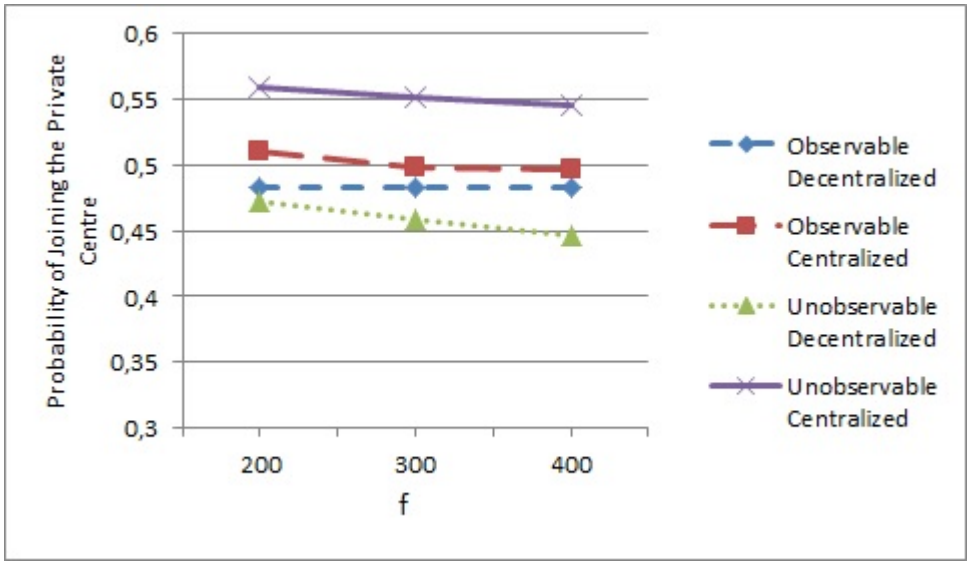

Figure 6.6: Sensitivity Analysis on Probability of Joining Private Centre with respect to f

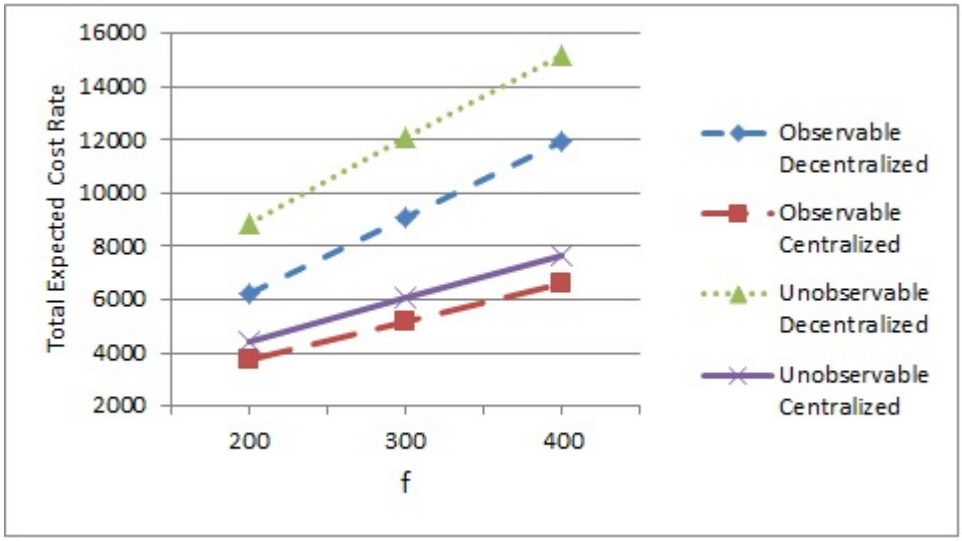

Figure 6.7: Sensitivity Analysis on Total Expected Cost with respect to f

As the private centre's fee increases, the probability of joining the private centre decreases (for the observable decentralized case, the amount of decreasing is too small.), while the probability of blocking and the total expected cost rate increase. According to Figure 6.6, changes in the value of f has a small effect on changing the probability of joining the private centre, especially for lower values of f. From Figure 6.7, the effect of f is more significant in the decentralized settings, when individuals choose to minimize their own expected cost.

Sensitivity Analysis with respect to c

We perform the sensitivity analysis for two values of c, 100 and 140.

Table 6.8: Performance Measures for $c = 100$

	Obs.Decent	Obs.Cent	Unobs.Decent	Unobs.Cent
Prob of joining private	0.483	0.497	0.458	0.549
Prob of Blocking	0.000	0.000	0.027	0.000
Total Expected Cost Rate	9083.09	5031.28	11551.392	5851.735

Table 6.9: Performance Measures for $c = 140$

	Obs.Decent	Obs.Cent	Unobs.Decent	Unobs.Cent
Prob of joining private	0.483	0.502	0.461	0.555
Prob of Blocking	0.000	0.000	0.025	0.000
Total Expected Cost Rate	9216.08	5308.52	12636.305	6269.64

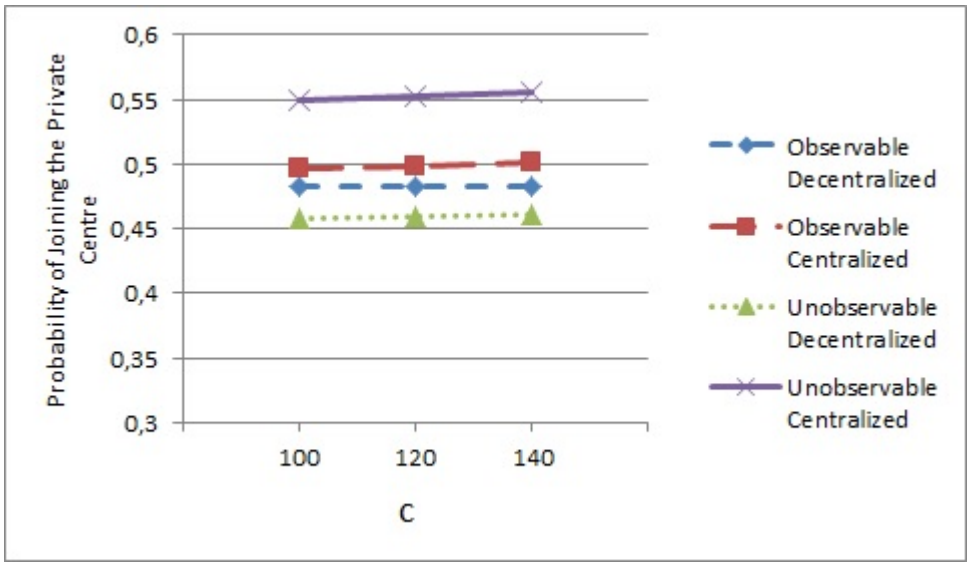

Figure 6.8: Sensitivity Analysis on Probability of Joining Private Centre with respect to c

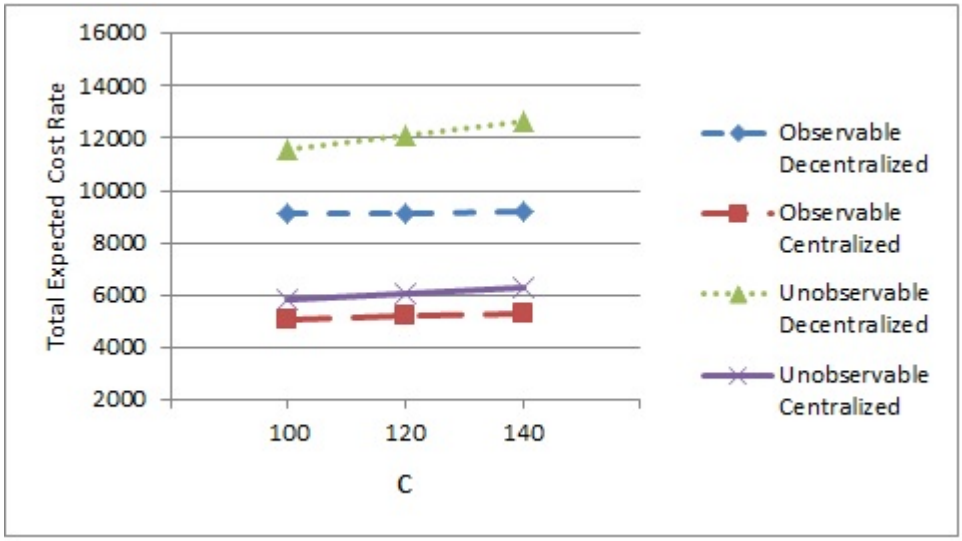

Figure 6.9: Sensitivity Analysis on Total Expected Cost with respect to c

As the waiting cost per unit of time increases, the probability of blocking, the total expected cost rate, and the probability of joining the private centre increase (for the observable decentralized case, the amount of increasing is too small.). According to Figure 6.8, the changes in the probability of joining the private centre with respect to c is too small. From Figure 6.9, the total expected cost also doesn't change significantly with respect to c. According to this figure, the total expected cost rate of the unobservable decentralized setting, which is the least optimal case, has the highest rate of change with respect to c. The lowest rate of change belongs to the observable centralized setting, which is the optimal case.

Sensitivity Analysis with respect to m_1

We perform the sensitivity analysis for two values of m_1, 40 and 60.

Table 6.10: Performance Measures for $m_1 = 40$

	Obs.Decent	Obs.Cent	Unobs.Decent	Unobs.Cent
Prob of joining private	0.483	0.498	0.459	0.554
Prob of Blocking	0.000	0.000	0.028	0.000
Total Expected Cost Rate	9082.32	5170.98	11422.543	6073.718

Table 6.11: Performance Measures for $m_1 = 60$

	Obs.Decent	Obs.Cent	Unobs.Decent	Unobs.Cent
Prob of joining private	0.483	0.498	0.46	0.552
Prob of Blocking	0.000	0.000	0.024	0.000
Total Expected Cost Rate	9084.91	5170.98	12785.453	6060.578

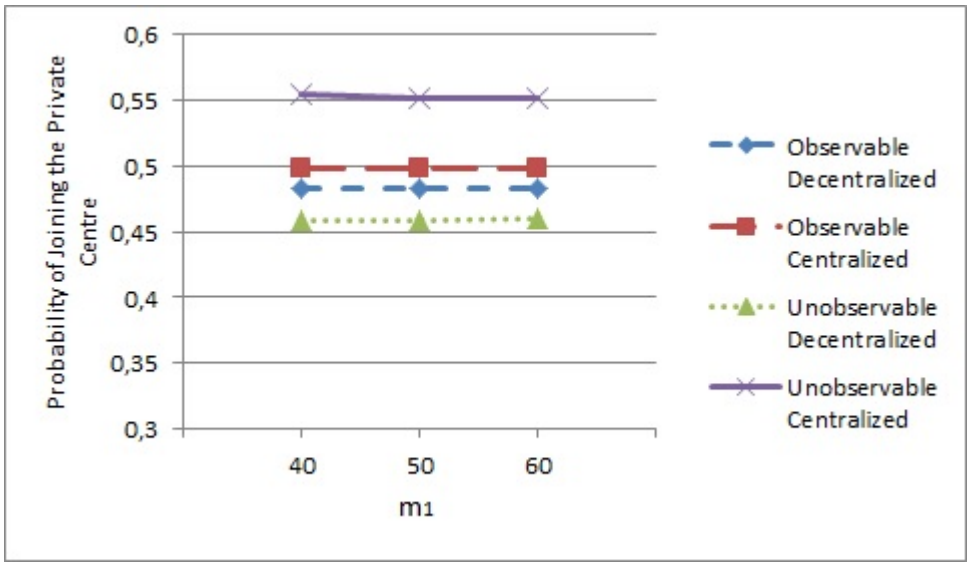

Figure 6.10: Sensitivity Analysis on Probability of Joining Private Centre with respect to m_1

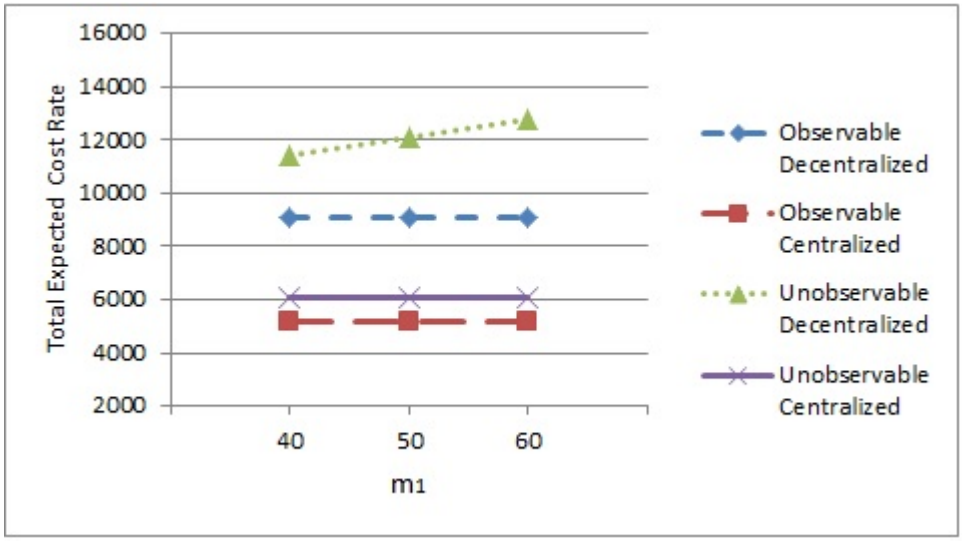

Figure 6.11: Sensitivity Analysis on Total Expected Cost with respect to m_1

As the buffer size of the free centre increases, the probability of blocking decreases. According to Figure 6.10, the probability of joining the private centre doesn't change considerably (It increases for all cases except the unobservable centralized case). According to Figure 6.11, the total expected cost rate is almost constant with respect to m_1 for all cases, except for the unobservable decentralized case. In this case, as we increase the buffer size of the free centre, the expected waiting time in that centre increases and leads to an increase in the total expected cost rate of the system.

Sensitivity Analysis with respect to c_B

We perform the sensitivity analysis for two values of c_B, 7500 and 10000.

Table 6.12: Performance Measures for $c_B = 7500$

	Obs.Decent	Obs.Cent	Unobs.Decent	Unobs.Cent
Prob of joining private	0.483	0.498	0.47	0.553
Prob of Blocking	0.000	0.000	0.018	0.000
Total Expected Cost Rate	9085.07	5170.98	11845.204	6065.996

Table 6.13: Performance Measures for $c_B = 10000$

	Obs.Decent	Obs.Cent	Unobs.Decent	Unobs.Cent
Prob of joining private	0.483	0.498	0.477	0.554
Prob of Blocking	0.000	0.000	0.014	0.000
Total Expected Cost Rate	9085.15	5170.98	11663.767	6069.074

As the probability of blocking is almost 0 in both observable cases, and also in the centralized unobservable case, the performance measures do not change considerably. In the unobservable decentralized setting, as the blocking cost increases, the probability of joining the private centre increases, the blocking probability and also the total expected cost rate decrease. The unobservable decentralized setting send a lower proportion of individuals to the private centre, in comparison to the optimal case (observable centralized). As the blocking cost increases, more patients tend to join the private centre to avoid the high cost of being blocked, so the probability of joining the private centre approaches to the optimal probability, and the total expected cost rate decreases.

Chapter 7

CONCLUSIONS

In this thesis, we analyse a healthcare queueing system of two parallel M/M/1 servers, one public and one private. Both queues have infinite buffer sizes. The individuals have to obtain service from either of the public or the private centre. We analyse the system in both decentralized and centralized settings. In the decentralized setting, individuals act in order to minimize their own expected cost, while in the centralized setting, there is a central authority who sends individuals to the servers in order to minimize the total expected cost of the system. We perform our analysis for both observable and unobservable queues.

We show that in the observable case, both in the decentralized and centralized settings, the optimal policy exists, and it is of a threshold type policy. We find a mathematical formula for calculating the thresholds in the observable decentralized case. The relationship between the thresholds in the decentralized and centralized setting is also examined. We prove that for each fixed number of patients in the public centre, the private thresholds are greater in the centralized setting, and for each fixed number of patients in the private centre, the public thresholds are greater in the decentralized setting. This implies that in the decentralized case, after some points, the arrivals who join the public centre have negative externalities on others.

In the unobservable case we establish the existence of a unique optimal strategy and show its structure in both decentralized and centralized cases.

We extend our results for the observable case to the model with finite buffer. For the unobservable case, we were not able to extend all results, but we show that the results are applicable under some conditions. Then, the sensitivity analysis is performed on some performance measures for all four cases with respect to different parameters of

the model.

This work can be extended in several directions. Some assumptions can be relaxed, for example we can consider balking and reneging with a positive probability. We can also relax the assumption that the service rate in the private centre is greater than the public centre and generalize the model. Because in some instances, the private server spend more time for each patient. It is also possible to assume different qualities for the centres and model the differences in qualities by considering different rewards of receiving service and different waiting costs for each centre. It can also be extended to two parallel M/M/m servers. We may consider different classes of customers with respect to their income, or prioritize the individuals with respect to their emergency levels.

The results of this study would direct the decision makers to decide on the ways they can improve the healthcare system with a limited budget. The effect of improving the public centre could be compared to the effect of assigning more subsidy for the private centre, based on the expenditure which is needed for each case. Moreover, the consequences of centralization and making the queues observable could be analysed and compared.

BIBLIOGRAPHY

[1] P. Naor, "The regulation of queue size by levying tolls," Econometrica: journal of the Econometric Society pp. 15–24 (1969).

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

[3] P. Musgrove, "Public and private roles in health: theory and financing patterns," (1996).

[4] E. Regidor, D. Martínez, M. E. Calle, P. Astasio, P. Ortega, and V. Domínguez, "Socioeconomic patterns in the use of public and private health services and equity in health care," BMC health services research **8**, 183 (2008).

[5] S. Basu, J. Andrews, S. Kishore, R. Panjabi, and D. Stuckler, "Comparative performance of private and public healthcare systems in low-and middle-income countries: a systematic review," PLoS med **9**, e1001244 (2012).

[6] C. Canta and M. L. Leroux, "Public and private health insurance, waiting times, and redistribution," CESifo Area Conference (2012).

[7] M. Jofre Bonet, "Public health care and private insurance demand: the waiting time as a link," Health Care Management Science **3**, 51–71 (2000).

[8] M. Johar, G. Jones, M. P. Keane, E. Savage, O. Stavrunova *et al.*, "The demand for private health insurance: Do waiting lists matter?–revisited," (2013).

[9] M. Johar, G. Jones, M. Keane, E. Savage, and O. Stavrunova, "Waiting times for elective surgery and the decision to buy private health insurance," Health Economics **20**, 68–86 (2011).

[10] B. Hajek, "Optimal control of two interacting service stations," Automatic Control, IEEE Transactions on **29**, 491–499 (1984).

[11] J. M. Harrison, "Dynamic scheduling of a multiclass queue: Discount optimality," Operations Research **23**, 270–282 (1975).

[12] J. C. Gittins, "Bandit processes and dynamic allocation indices," Journal of the Royal Statistical Society. Series B (Methodological) pp. 148–177 (1979).

[13] Z. Rosberg, P. Varaiya, and J. C. Walrand, "Optimal allocation of resources between research projects," IEEE Transactions on Automatic Control **27**, 600–609 (1982).

[14] P. Whittle *et al.*, "Arm-acquiring bandits," The Annals of Probability **9**, 284–292 (1981).

[15] P. Varaiya, J. Walrand, and C. Buyukkoc, "Extension of the multi-armed bandit problem," pp. 1179–1180 (1983).

[16] W. Lin and P. Kumar, "Stochastic control of a queue with two servers of different rates," Analysis and Optimization of Systems pp. 719–728 (1982).

[17] M. J. Sobel, "The optimality of full service policies," Operations Research **30**, 636–649 (1982).

[18] W. Winston, "Optimality of the shortest line discipline," Journal of Applied Probability pp. 181–189 (1977).

[19] R. R. Weber, "On the optimal assignment of customers to parallel servers," Journal of Applied Probability pp. 406–413 (1978).

[20] I. Adler and P. Naor, "Social optimization versus self-optimization in waiting lines," (1969).

[21] W. Lin and P. Kumar, "Optimal control of a queueing system with two heterogeneous servers," Automatic Control, IEEE Transactions on **29**, 696–703 (1984).

[22] A. Burnetas and I. Georgiou, "Customer equilibrium and optimal strategies in two unobservable queues," Manuscript in Preparation (2015).

[23] V. A. Knight and P. R. Harper, "Selfish routing in public services," European Journal of Operational Research **230**, 122–132 (2013).

[24] P. Guo, R. Lindsey, and Z. G. Zhang, "On the downs-thomson paradox in a self-financing two-tier queuing system," Manufacturing & Service Operations Management **16**, 315–322 (2014).

[25] J. E. Cohen and F. P. Kelly, "A paradox of congestion in a queuing network," Journal of Applied Probability pp. 730–734 (1990).

[26] R. Hassin and R. Roet, "Equilibrium in a two dimensional queueing game: When inspecting the queue is costly," (2011).

[27] A. Ephremides, P. Varaiya, and J. Walrand, "A simple dynamic routing problem," Automatic Control, IEEE Transactions on **25**, 690–693 (1980).

[28] K. S. Anand, M. F. Pac, and S. Veeraraghavan, "Quality-speed conundrum: trade-offs in customer-intensive services," Management Science **57**, 40–56 (2011).

[29] L. V. Green, S. Savin, and B. Wang, "Managing patient service in a diagnostic medical facility," Operations Research **54**, 11–25 (2006).

[30] S. H. Xu, R. Righter, and J. G. Shanthikumar, "Optimal dynamic assignment of customers to heterogeneous servers in parallel," Operations Research **40**, 1126–1138 (1992).

[31] J. Suk and C. G. Cassandras, "Optimal scheduling of two competing queues with blocking," pp. 1102–1107 (1988).

[32] J. Baras, A. Dorsey, and A. Makowski, "Two competing queues with linear costs and geometric service requirements: The μc-rule is often optimal," Advances in Applied Probability pp. 186–209 (1985).

[33] J. Wang, O. Baron, and A. Scheller-Wolf, "M/m/c queue with two priority classes," Operations Research (2015).

[34] R. B. Myerson, *Game theory* (Harvard university press, 2013).

[35] J. Resing and L. Örmeci, "A tandem queueing model with coupled processors," Operations Research Letters **31**, 383–389 (2003).

[36] D. P. Bertsekas, *Dynamic programming and optimal control*, vol. 1 (Athena Scientific Belmont, MA, 1995).

[37] M. L. Puterman, *Markov decision processes: discrete stochastic dynamic programming* (John Wiley & Sons, 2014).

[38] S. M. Ross, *Introduction to probability models* (Academic press, 2014).

Appendix

Steady-State Equations for the Observable Decentralized Case with Infinite Buffer

$$\lambda\pi_{i,j} = \mu_1\pi_{i+1,j} + \mu_2\pi_{i,j+1} \quad \text{if } i = j = 0$$

$$(\lambda + \mu_2)\pi_{i,j} = \mu_2\pi_{i,j+1} + \mu_1\pi_{i+1,j} \quad \text{if } i = 0, j \geq 1$$

$$(\lambda + \mu_1)\pi_{i,j} = \lambda\pi_{i-1,j} + \mu_1\pi_{i+1,j} + \mu_2\pi_{i,j+1} \quad \text{if } 1 \leq i \leq n_1^1, j = 0$$

$$(\mu_1 + \lambda)\pi_{i,j} = \mu_1\pi_{i+1,j} + \mu_2\pi_{i,j+1} \quad \text{if } i > n_1^1, j = 0$$

$$(\lambda + \mu_1 + \mu_2)\pi_{i,j} = \lambda\pi_{i-1,j} + \lambda\pi_{i,j-1} + \mu_2\pi_{i,j+1} + \mu_1\pi_{i+1,j} \quad \text{if } i = n_1^k, n_2^{k-1} < j < n_2^k$$

$$(\lambda + \mu_1 + \mu_2)\pi_{i,j} = \lambda\pi_{i,j-1} + \mu_1\pi_{i+1,j} + \mu_2\pi_{i,j+1} \quad \text{if } i > n_1^k, n_2^{k-1} < j < n_2^k$$

$$(\lambda + \mu_1 + \mu_2)\pi_{i,j} = \lambda\pi_{i-1,j} + \mu_1\pi_{i+1,j} + \mu_2\pi_{i,j+1} \quad \text{if } 1 \leq i < n_1^{k+1} - 1, j = n_2^k$$

$$(\lambda + \mu_1 + \mu_2)\pi_{i,j} = \lambda\pi_{i,j-1} + \lambda\pi_{i-1,j} + \mu_1\pi_{i+1,j} + \mu_2\pi_{i,j+1} \quad \text{if } i = n_1^{k+1} - 1, j = n_2^k$$

$$(\lambda + \mu_1 + \mu_2)\pi_{i,j} = \lambda\pi_{i-1,j} + \mu_2\pi_{i,j+1} + \lambda\pi_{i,j-1} + \mu_1\pi_{i+1,j} \quad \text{if } i = n_1^{k+1}, j = n_2^k$$

$$(\lambda + \mu_1 + \mu_2)\pi_{i,j} = \lambda\pi_{i,j-1} + \mu_2\pi_{i,j+1} + \mu_1\pi_{i+1,j} \quad \text{if } i > n_1^{k+1}, j = n_2^k$$

$$\sum_{j=0}^{\infty}\sum_{i=0}^{\infty}\pi_{i,j} = 1$$

Steady-State Equations for the Observable Decentralized Case with Finite Buffer

$$\lambda\pi_{i,j} = \mu_1\pi_{i+1,j} + \mu_2\pi_{i,j+1} \quad \text{if } i = j = 0$$

$$(\lambda + \mu_2)\pi_{i,j} = \mu_1\pi_{i+1,j} + \mu_2\pi_{i,j+1} \quad \text{if } i = 0, 1 \le j < m_2$$

$$(\lambda + \mu_2)\pi_{i,j} = \mu_1\pi_{i+1,j} \quad \text{if } i = 0, j = m_2$$

$$(\lambda + \mu_1)\pi_{i,j} = \lambda\pi_{i-1,j} + \mu_1\pi_{i+1,j} + \mu_2\pi_{i,j+1} \quad \text{if } 1 \le i \le n_1^1, i < m_1, j = 0$$

$$(\lambda + \mu_1)\pi_{i,j} = \lambda\pi_{i-1,j} + \mu_2\pi_{i,j+1} \quad \text{if } 1 \le i = m_1 \le n_1^1, j = 0$$

$$(\lambda + \mu_1)\pi_{i,j} = \mu_1\pi_{i+1,j} + \mu_2\pi_{i,j+1} \quad \text{if } n_1^1 < i < m_1, j = 0$$

$$(\lambda + \mu_1 + \mu_2)\pi_{i,j} = \lambda\pi_{i-1,j} + \mu_1\pi_{i+1,j} + \mu_2\pi_{i,j+1}$$
$$\text{if } 1 \le i < n_1^k \le m_1, n_2^{k-1} < j < n_2^k \le m_2$$
$$\text{or } 1 \le i < n_1^{k+1} - 1 \le m_1, j = n_2^k < m_2$$

$$(\lambda + \mu_1 + \mu_2)\pi_{i,j} = \lambda\pi_{i-1,j} + \mu_1\pi_{i+1,j}$$
$$\text{if } 1 \le i < n_1^k < m_1, n_2^{k-1} < j = m_2 < n_2^k$$
$$\text{or } 1 \le i < n_1^{k+1} - 1 \le m_1, j = n_2^k = m_2$$
$$\text{or } i = n_1^{k+1} - 1 = m_1 - 1, j = n_2^k = m_2, m_2 = n_2^{m_1 - n_1^1 + 1}$$

$$(\lambda + \mu_1 + \mu_2)\pi_{i,j} = \lambda\pi_{i-1,j} + \lambda\pi_{i,j-1} + \mu_1\pi_{i+1,j} + \mu_2\pi_{i,j+1} \quad \text{if } i = n_1^k < m_1, n_2^{k-1} < j < n_2^k \le m_2$$

$$(\lambda + \mu_1 + \mu_2)\pi_{i,j} = \lambda\pi_{i,j-1} + \mu_1\pi_{i+1,j} + \mu_2\pi_{i,j+1} \quad \text{if } n_1^k < i < m_1, n_2^{k-1} \le j < n_2^k \le m_2$$

$$(\lambda + \mu_1 + \mu_2)\pi_{i,j} = \lambda\pi_{i-1,j} + \lambda\pi_{i,j-1} + \mu_1\pi_{i+1,j}$$
$$\text{if } n_1^k \le i < m_1, n_2^{k-1} < j = m_2 < n_2^k$$
$$\text{or } n_1^{k+1} \le i < m_1, j = n_2^k = m_2$$
$$\text{or } i = n_1^{k+1} - 1 < m_1 - 1, j = n_2^k = m_2$$
$$\text{or } i = n_1^{k+1} - 1 = m_1 - 1, j = n_2^k = m_2, m_2 \ne n_2^{m_1 - n_1^1 + 1}$$

$$(\lambda + \mu_1 + \mu_2)\pi_{i,j} = \lambda\pi_{i-1,j} + \lambda\pi_{i,j-1} + \mu_1\pi_{i+1,j} + \mu_2\pi_{i,j+1}$$
$$\text{if } i = n_1^{k+1} < m_1, n_2^k < j \le n_2^{k+1} < m_2$$
$$\text{or } i = n_1^{k+1} - 1 < m_1, j = n_2^k < m_2$$

$$(\lambda + \mu_1)\pi_{i,j} = \mu_2\pi_{i,j+1} \quad \text{if } i = m_1 \ne n_1^1, j = 0$$

$$(\lambda + \mu_1 + \mu_2)\pi_{i,j} = \lambda\pi_{i,j-1} + \mu_2\pi_{i,j+1}$$
$$\text{if } i = m_1, n_2^{k-1} < j \le n_2^k, j < m_2, j < n_2^{m_1 - n_1^1 + 1}$$

$$(\lambda + \mu_1 + \mu_2)\pi_{i,j} = \lambda\pi_{i-1,j} + \lambda\pi_{i,j-1} + \mu_2\pi_{i,j+1} \quad \text{if } i = n_1^k = m_1, n_2^{k-1} < j \le n_2^k, j < m_2$$

$$(\mu_1 + \mu_2)\pi_{i,j} = \lambda\pi_{i-1,j} + \lambda\pi_{i,j-1} \quad \text{if } i = m_1, j = m_2$$

$$\sum_{j=0}^{m_2}\sum_{i=0}^{m_1}\pi_{i,j} = 1$$