Analysis of On-Demand Ride-Hailing Systems

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Guiyun Feng

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Advisors: Prof. Zizhuo Wang, Prof. Guangwen Kong

October, 2018

Acknowledgements

Firstly, I would like to express my sincere gratitude to my advisor Prof. Zizhuo Wang. It has been my great honor to be his Ph.D. student. In the past four years he has shown me, by his brilliant creativity and passionate attitude at work, and his kindness and generosity to people around, what a great researcher and person he is. He meets with students regularly and is always available to us. He reads drafts promptly and provides clear, constructive comments. He sees the strengths and skills of each individual and then designs projects to suit these strengths. His continued passion and complete absorption in academic research stimulate me to be persistent when encountering difficulties along the journey. His advice on both research as well as on my career have been invaluable. I would always remember his effort in introducing me to the area of Revenue Management and teaching me how to conduct scientific research, his prompt response to any difficulty I encountered, his help in revising my notes and polishing my presentation skills. I cannot be more thankful for his tremendous support along my job seeking process. I deeply appreciate all his contributions of time, ideas, and funding to make my Ph.D. experience productive and stimulating.

My special gratitude also goes to my co-advisor Prof. Guangwen Kong. Without the occasional talk with Prof. Kong in the early 2014 and her encouragement, I may not have started my fruitful journey in the University of Minnesota. I feel lucky to work with her

closely on a very interesting topic. Her expertise in Operations Management inspires me to think in a managerial point of view. Her patience, support and consideration make the research collaboration very enjoyable. The passion and joy she has for her research was inspiring and I am thankful for the excellent example she has provided as a successful female faculty and a great mother.

Besides my advisors, I would like to thank the rest of my dissertation committee members: Prof. Saif Benjaafar, Prof. William L. Cooper and Prof. Yousef Saad, for their insightful comments and encouragement, as well as the hard questions which motivate me to widen my research from various perspectives. I am quite appreciative of Prof. Saad for taking time from his busy traveling schedule to serve on my thesis committee. I wish to thank Prof. Benjaafar for teaching me two excellent and fundamental courses in Supply Chain Management. I am very thankful for the tremendous support from Prof. Benjaafar and Prof. Cooper when I was on the job market last year: thank you sincerely for the academic references and all effort in improving my interview and presentation skills.

I thank my fellows in UMN: Rui Chen, Xiang Li, Hamidreza Badri, Behrooz Pourghannad, Junfeng Zhu, Xiao Chen, Yuanchen Su, Jiali Huang, Xiang Gao, Shaozhe Tao, Ruizhi Shi, Zeyang Wu, Yiru Wang, Meng Zhou and Wen Xing, for the stimulating discussions in research and life, and for all the fun we have had in the last four years. My time at UMN was made enjoyable in large part due to the many friends. Also, my gratitude goes to my old friends for sticking by me through ups and downs: Ying Huang, Shuhui Wu, Xiaohan Shen, Yanchao Jiang, Ying Zhao, Danru Qu, Na Xu, Xuan Zhao and Shan Wu. Thank you for making me feel loved and appreciated.

Lastly, I would like to thank my parents for giving me life, raising me up with plenty of effort, love, and encouragement, providing me all the freedom and support to pursue my dreams. I must express my gratitude to my soulmate and my husband, Xiaobo Li.

His gentle, thoughtful and loving mind is the source of strength, peace, and happiness in my life.

Dedication

To the memory of my beloved father.

Abstract

Recently, there has been a rapid rise of on-demand ride-hailing platforms, such as Uber and Didi, which allow passengers with smartphones to submit trip requests and match them to drivers based on their locations and drivers availability. This increased demand has raised questions about how such a new matching mechanism will affect the efficiency of a transportation system, in particular, whether it will help reduce passengers average waiting time compared to traditional street-hailing systems. In this dissertation, we address this question by building a stylized model of a circular road and comparing the average waiting times of passengers under various matching mechanisms. After identifying key tradeoffs between different mechanisms, we find that surprisingly, the on-demand matching mechanism could result in higher or lower efficiency than the traditional street-hailing mechanism, depending on the parameters of the system. To overcome the disadvantage of both systems, we further propose adding response caps to the on-demand hailing mechanism and develop a heuristic method to calculate a near-optimal cap. We also test our model using more complex road networks and examine the impact of passenger abandonments, idle time strategies of taxis and traffic congestion on the performance of the ride-hailing systems. The results of this research would be instrumental for understanding the tradeoffs of the new service paradigm and thus enable policy makers to make more informed decisions when enacting regulations for this emerging service paradigm.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In the past few years, with the growth of information technology, we have witnessed the rapid rise of on-demand ride-hailing platforms such as Uber, Lyft and Didi. Thanks to these platforms, passengers nowadays can request rides using their smartphones instead of hailing taxis on the streets, which used to be the norm in big metropolitan areas. Ride-hailing platforms help connect passengers with drivers (private drivers and traditional taxi drivers) in real time. The growth of these platforms has been astonishing: Uber completed its two-billionth trip in July 2016 (Hawkins 2016b), Didi was responsible for 1.43 billion trips in 2015 alone (Hawkins 2016a), and the growth continues to accelerate.

Despite the rapid growth, it has been debated about whether or to what extent such platforms should be encouraged by policy makers worldwide. Prominent in the debate are concerns about safety, privacy and liability of the drivers and the platform (Rogers 2015). Yet there is also the central question of whether these platforms always increase convenience for passengers and the efficiency of the transportation system. In particular, policy makers struggle to understand how much increased efficiency, if any, such platforms bring to the transportation system; whether the platform can reduce passengers' average waiting times when they need a ride, and if the platform has negative

effects on the transportation system. Understanding the answers to these questions would enable policy makers to make more informed decisions when enacting regulations for this emerging service paradigm.

Unfortunately, the answers to these questions are not obvious. Even though it is generally believed that such platforms may help passengers obtain transportation more efficiently in some circumstances, it is also realized that the efficiency gained through on-demand matching may disappear in others, especially when the supply could easily match the demand without the use of ride-hailing platforms. For example, Steinberg (2012) points out that "People waiting on Manhattan's Fifth Avenue in the middle of rush hour don't need an app to tell cabs to come to the area." Moreover, a recent consulting report by Schaller (2017) identifies a phenomenon called "empty seats, busy street." Based on the data collected in Manhattan, New York, Schaller (2017) finds that the average unoccupied time between trips for app-based ride services such as Uber and Lyft is 11 minutes, which is significantly longer than the average 8 minutes for yellow cabs that provide street-hailing services. It has been suspected that a driver heading to pick up a passenger who requested service from a platform is likely to encounter another passenger en route in a high-traffic situation, thus increasing the otherwise shorter pickup time and resulting in a loss of efficiency. An example of actions taken in response to the potential inefficiency of an on-demand ride-hailing platform is the Shanghai government's 2014 ban on the use of on-demand ride-hailing platforms during peak traffic hours (Sweeney 2014). Although the ban was later lifted, many questions remain regarding the performance of an on-demand ride-hailing platform compared to a street-hailing system.

This research attempts to shed light on the performance of an on-demand ride-hailing system from an operations point of view. We aim to evaluate the on-demand ride-hailing system under different circumstances, and to suggest ways to address the inefficiency

arising from the on-demand ride-hailing matching mechanism. Specifically, we consider a model in which taxis[1] drive on a circular road, and we study the performance of different matching mechanisms for the proposed model. This model captures the main tradeoffs between the street-hailing system and the on-demand hailing system and thus provides useful insights. In particular, for the street-hailing system, we consider a corresponding "no-call mechanism" in which no platform matches passengers with taxis, and arriving passengers are picked up by the first available taxi passing by. For the on-demand hailing system, we consider a corresponding "call mechanism" in which waiting passengers are matched to the nearest idle taxi whenever there are taxis available. We find that the on-demand ride-hailing system is different from street-hailing in that the average waiting time is non-monotone in the system utilization level. Such non-monotonicity is because of the non-monotonicity of the average en route time in such systems, which captures the time between a matching is established and the actual pickup occurs. By comparing the on-demand ride-hailing and street-hailing systems, we identify the advantages and disadvantages of each mechanism. The advantage of the call mechanism lies in its ability to tell the driver where (which direction) to go to pick up the next passenger. However, taxis using the call mechanism may suffer from the possibility of forgoing future better matching opportunities when accepting an incoming request. Upon understanding these tradeoffs, we illustrate the conditions under which one mechanism is more efficient than the other. In a setting where passengers do not abandon the system before being served, we show that the call mechanism is more efficient in terms of passengers waiting time when the traffic intensity is either low or high, while the no-call mechanism could be more efficient when the traffic intensity is in the middle range. Similar conclusion has also been found in a more complex transportation network model — a grid network, where

[1]In this dissertation, we use "taxis" to refer to the cars that are used to transport passengers. The term may include traditional taxis, as well as private cars or even self-driving cars that are used to provide transportation services.

the no-call mechanism outperforms the call mechanism at medium system utilization level when the density of taxis is high and the grid network itself is not too complex.

Based on the tradeoffs between the call and no-call mechanisms, we further propose a modified call mechanism that adds a distance cap such that a passenger is matched to a taxi only when their distance is below the cap. We show that adding such a cap would preserve most of the benefit of the call mechanism, while greatly diminishing the disadvantage, thus achieving a superior performance. We further propose a heuristic method to calculate a near-optimal cap that is shown to provide good performance compared to the optimal selection.

We also examine the performance of the call and no-call matching mechanisms in various other settings. First, we consider the setting where passengers may abandon the system before being served. In this setting, we find that the call mechanism might be less efficient compared to the no-call mechanism when the passenger's patience level is low and the traffic intensity is high. More specifically, in the call mechanism, fewer passengers would be served and the served passengers would have a longer waiting time, which is consistent with the empirical findings in Schaller (2017). Then we examine the impact of different idle time strategies and traffic congestion on the performance of the call and no-call mechanisms. We believe that these extensions could be instrumental for decision makers to understand the tradeoffs of the new service paradigm in different settings.

1.1 Organization

The remainder of the dissertation is organized as follows. In Chapter 2, we review the literature that is related to the dissertation. In Chapter 3, we first introduce the transportation network model, as well as the call and no-call mechanisms in Section

3.1. In Section 3.2, we illustrate the service performance for both call and no-call mechanisms by simulation experiments, and further provide insights to understand the results. In Section 3.3, we develop approximation schemes for both systems, which enable us to validate the important observations from Section 3.2. In Section 3.4, we compare the efficiency between the call and no-call systems, and obtain our main results. In Chapter 4, we propose a distance cap to add to the call mechanism and study the optimal selection of the cap. Numerical experiments on more complex road networks are conducted in Chapter 5. The impact of several important factors are discussed in Chapter 6. We conclude the dissertation by summarizing its contribution and future directions to explore in Chapter 7. All proofs are relegated to the Appendix A.

Chapter 2

Literature Review

The emergence and popularity of the sharing economy in recent years have raised many interesting research questions and attracted significant academic interest. The research conducted in this area can be categorized into strategic and operational levels. In the following, we review literature in both categories.

On the strategic level, several recent works examine how the emergence of the sharing economy changes the way people behave and subsequently its impact on the economy. For example, Fraiberger and Sundararajan (2017) use aggregate data from the U.S. automobile industry to show that a shift from ownership to collaborative consumption may lead to less usage, lower used-good prices and a higher consumer surplus. Benjaafar et al. (2018b) consider an equilibrium model that endogenizes market friction and provide analytical results showing that product usage and ownership may increase with the presence of a sharing platform. Jiang and Tian (2016) examine the impact of a product-sharing platform on the manufacturer's profit and consumers' surplus. They find that when product quality is exogenous, the manufacturer and consumers are better off when the marginal production cost is high, but both are worse off otherwise. However, when the product quality is endogenous, the consumer surplus is always lower

in the presence of a sharing platform. In comparison to the research focusing on product sharing, our study considers operational decisions in matching supply and demand of transportation service via on-demand hailing platforms and studies the impact of different matching mechanisms. In addition, most papers consider consumer surplus that depends on product price and quality. Instead, we focus on whether an on-demand hailing platform can reduce average passenger waiting time.

Another growing stream of papers focus on the operational decision-making problems faced by the on-demand service platforms. One problem that has been considered extensively is capacity management via dynamic pricing. In particular, on-demand service platforms have enabled service providers to choose their own flexible work schedules, presenting new challenges to managing the service capacity. To address this problem, analysis often focuses on how the platform can adjust agent payment and service price to efficiently allocate capacity. For example, Gurvich et al. (2015) employ a newsvendor model to study the capacity management problem in sharing marketplaces where workers have the flexibility to choose their own work schedules. Riquelme et al. (2015) model a ride-sharing platform as a queue with customers' arrival and the drivers' work hours depending on the real-time dynamic service price, and they show that the platform cannot significantly increase its revenue by using a dynamic pricing policy based on a threshold number of drivers. Cachon et al. (2017) consider several contractual forms ranging from fixed price/compensation to surge pricing under a two-period framework and find that providers and consumers are generally better off with surge pricing. Taylor (2016) examines how two defining features of an on-demand service platform – congestion-driven delay disutility and agent independence – affect the platform's optimal per-service price and wage. Guda and Subramanian (2018) study the role of surge pricing in managing work availability across market locations and they show that, contrary to conventional wisdom, surge pricing can be useful for locations with excess

demand by motivating more drivers moving to other locations. Bimpikis et al. (2016) provide the impact of demand pattern across locations on the platform's prices, profit and consumer surplus. Castillo et al. (2017) describe the Wild Goose Chase (WGC) phenomenon caused by too thinly spread idle taxis throughout a city and two pricing schemes are provided to avoid WGCs. Benjaafar et al. (2017) specify how car ownership and congestion are effected by platform prices, cost parameters, and the distribution of the individuals' types. They discover that the traffic and ownership may increase as the ownership cost increases, and that a revenue maximizing platform might prefer a situation where cars are driven with only a few seats occupied, creating high congestion. Benjaafar et al. (2018a) study labor welfare in on-demand service platforms and show that factors that affect labor supply, such as labor pool size, delay cost, and variability in the agents' opportunity cost, may have a non-monotonic effect on labor welfare. Castro et al. (2018) characterize the optimal pricing policy and the implications of the strategic nature of supply units. In particular, they show that the platform will use prices to create artificially damaged regions where driver congestion is artificially high in order to lure drivers towards more profitable locations for the platform. Hall et al. (2017) estimate the effects of sudden fare changes on market outcomes and their result show the driver supply of labor to ride-sharing markets is highly elastic. Tang et al. (2016) use the steady-state equilibrium to characterize the optimal price, wage and pay-out ratio that maximize the profit of the platform, where an $M/M/1$ queuing model is used to get the approximated waiting time for passengers. In our study, we do not focus on the pricing and capacity decisions. Instead, we assume the demand and the supply are exogenously given and focus on comparing different matching mechanisms under various utilization levels.

In addition to capacity management, another important operational problem faced by platforms is how to efficiently match service providers with customers. Several recent

works have taken on this task. Allon et al. (2012) explore the role of the platforms in facilitating information gathering, operational efficiency and communication by considering three different market models employed by platforms and then characterize the corresponding market outcomes. Cullen and Farronato (2014) study the problem of balancing highly variable demand and supply for a frictional matching market and calibrate their model by using data from TaskRabbit. Anderson et al. (2015) investigate timely exchanges for agents in a barter marketplace, and their results show that a greedy policy that attempts to match upon each arrival is approximately optimal (minimizes average waiting time) among a large class of policies including batching policies. Baccara et al. (2015) consider two-sided matching with vertically different preferences over agents. They show that the optimal mechanism always matches congruent pairs immediately and holds on to a stock of incongruent pairs up to a certain threshold, and a centralized market is more appealing than the corresponding decentralized one. Akbarpour et al. (2016) study different dynamic matching strategies in network markets where agents arrive stochastically and stay for a random period of time before leaving the system if not matched. They show that waiting to thicken the market is highly valuable if the central planner knows the agents' departure time, otherwise, a greedy local algorithm is close to optimal. Hu and Zhou (2016) model dynamic matching between the demand and supply of heterogeneous types in a periodic-review fashion. They provide sufficient conditions on matching rewards such that the optimal matching policy follows a priority hierarchy among possible matching pairs. Similar to these works, the on-demand matching problem studied in our work is also a two-sided dynamic matching problem faced by a centralized platform; in particular, we can model the preference level (matching quality) for any taxi-passenger pair by their distance. However, it would be challenging to use the methods in these papers to evaluate the efficiency of the system. This is because a key differentiating factor in our study is that since taxis keep moving, the

matching scores are constantly evolving even when no new arrival occurs. We obtain insights that are particular to this dynamic.

In addition to the operations management literature, the dynamic matching problem described above has also been studied in the transportation literature under the taxi fleet dispatch context. Meyer and Wolfe (1961) analyze dispatching strategies and idle-time strategies when demand arises in two fixed location and served by one taxi, and extends the analysis to operations of a taxi fleet in a two dimensional region using an approximation of the stationary state. Bailey and Clark (1992) present a simulation model to study the average waiting time and taxi utilization of an urban taxi system under different dispatch and idle-time strategies. Gerrard (1974) compare the difference of operations between a a dial-a-b bus services and taxi affected, and examine how they are affected by the change of demand. The above papers do not consider street hailing where a taxi could actively search passengers without being dispatching by a centralized system. Instead, our work develops a unified frame work to compare two matching strategies: street-hailing and on-demand hailing, and we obtain insights about under what circumstances each strategy has better performance. The one with the most extensive contribution to date is McLeod (1972). McLeod (1972) develops a queueing approximation for the taxi system and propose several possible matching mechanisms. In addition, McLeod (1972) also acknowledges that the matching mechanism would play a critical role in en route time. However, the queueing approximation in McLeod (1972) does not reflect the difference of customer waiting time under different matching mechanism. Under the approximation in McLeod (1972), customer waiting time always increases with system work load. To the best of our knowledge, we are the first to show the non-monotonicity of customer waiting time in system work load. More recently, Ozkan and Ward (2017) propose a CLP-Based matching policy that improves the closest driver (CD) policy when arrival rates are heterogeneous in different region. Afèche

et al. (2018) considers self-interested drivers strategically make reposition decisions when demand arrivals are imbalanced in two locations. We propose potential ways to eliminate the disadvantage of both strategies and thus improve the efficiency of the system when arrivals distributed uniformly in the system.

Our study is also related to various queueing models and optimization models in vehicle routing. The queueing model that is closest to our problem is M/G/K queue with heterogeneous servers (Boxma et al. 2002, Sani and Daman 2015, Gupta et al. 2010). This is because the service time in our model could be decomposed two parts: the effective service time that follows an exponential distribution and the en route time that is depends on the number of taxis and waiting customers in the system. Boxma et al. (2002) provide an asymptotic analysis on the M/G/2 queuing system where the service time with a server follows an exponential distribution and that with another follows a general distribution. Sani and Daman (2015) derive the steady state distribution for the number of customers in the system and the mean waiting time in such a system. However, the same analysis is difficult to carry over to a system with more than two servers. Although M/G/K queueing system is one of the oldest model in queueing literature, the approximation of such system studied in literature are not well performed especially when the variance of the service time or the system utilization is high. Gupta et al. (2010) show the inapproximation of M/G/K queuing system using first two moments to approximate mean waiting time. Note that although it is appealing to resort to heavy traffic approximation in queueing to resolve our problem, the impact of en-route time under different matching mechanisms disappears when the system load approaches to 1. Our model is also related to polling system (Coffman and Gilbert 1986, Kroese and Schmidt 1992) where there are multiple queues and each of the servers visits these queues according to its own cyclic schedule. The difference of problems to be solved is that in our street-hailing service model, taxis have no cyclic schedule of

visiting the queues, and they simply pick up the first waiting passenger encountered. In addition, our problem also has the connection to assignment problem in optimization in both static setting (e.g., see Dantzig 2016, Murty 1992, Balinski and Gomory 1964, Tomizawa 1971, Jonker and Volgenant 1987, Barr et al. 1977, Hung 1983, Bertsekas 1981) and dynamic settings(e.g., see Pinedo 2016, Shmoys et al. 1995, Hall et al. 1997, Hoogeveen and Vestjens 2000, Spivey and Powell 2004). These research papers focus on developing optimization algorithms. In contrast to those research, we highlight on how the inefficiency arises when using such a dispatching system, and propose a simple yet implementable policy to avoid inefficiency.

Chapter 3

Models and Analysis

In this chapter, we introduce the stylized matching models for street-hailing and on-demand ride-hailing platforms, compare their efficiency and provide insights for the performance difference under different transportation parameters.

3.1 Model

We consider a stylized transportation system on a circular road with perimeter R. There are k taxis in the system. Passengers arrive at the system according to a Poisson process with rate λ, and their arrival locations are uniformly distributed on the circle. Each arriving passenger requests a service with distance D following an exponential distribution with mean d.[1] We assume that all taxis have a constant speed of v; i.e., they travel v units of distance on the circle per unit of time. Hence, the service duration follows an exponential distribution with mean d/v. For notational ease, we further define the service rate as $\mu = v/d$, and the utilization level as $\rho \doteq \lambda/k\mu$, which is an

[1] In our model, it is possible for the travel distance of a passenger to be greater than R. However, our model can be easily modified to one in which the requested travel distances of the passengers are bounded by R.

indicator of the traffic intensity of the system. In this dissertation, we consider two different systems in terms of the travel direction: the one-direction system and the two-direction system. In the one-direction system, all taxis travel clockwise, and so are the passengers' requests (i.e., if a passenger requests a service with distance l, then she is requesting to be transported in a clockwise direction for a distance of l). In contrast, in the two-direction system, taxis and passengers are allowed to travel in both directions. More precisely, the initial direction of each taxi, the requested service direction of each passenger, as well as the travel direction of a taxi after completing a service in the two-direction system, are all randomly chosen, with a 50% chance being clockwise and a 50% chance being counterclockwise.

The focus of this dissertation is on comparing two supply-demand matching mechanisms. The first mechanism is referred to as the *call mechanism*, which is used to model the matching process made by the on-demand ride-hailing platforms. In the call mechanism, when a passenger arrives, a platform immediately matches the passenger to the nearest taxi if there is a taxi available; otherwise, arrived passengers will wait until the next taxi becomes available, at which time the taxi will be matched to its nearest passenger (the remaining passengers, if any, would keep waiting for the next taxi to become available). Once a match between a taxi and a passenger is made, the taxi is committed to serving that passenger next, and the passenger waits for the matched taxi to come. We will consider a case in which passengers can only wait up to a certain amount of time in Section 6.1. In the one-direction system, the distance between a passenger and a taxi is defined as the distance for a taxi to reach a passenger when driving clockwise; while in the two-direction system, the distance is defined as the shorter distance between driving clockwise and counter-clockwise. The other mechanism is the *no-call mechanism*, which is used to model the matching process of traditional street-hailing. In the no-call mechanism, no matching occurs, and a passenger is picked

up by the first available taxi passing by.

To summarize, we consider four settings in this dissertation: one-direction call/no-call systems and two-direction call/no-call systems. The features of the four settings are listed in Table 3.1.

Table 3.1: Features of call/no-call mechanisms in one-direction/two-direction systems

Transportation systems	*1-d no-call*	*1-d call*	*2-d no-call*	*2-d call*
Driving direction	Clockwise	Clockwise	Both	Both
Initial direction/ Request direction/ Direction upon finish	Clockwise	Clockwise	Random	Random
Matching mechanism	No matching	Matched to nearest clockwise	No matching	Matched to nearest between two directions

The goal of this work is to study and compare the efficiency of these systems, particularly call versus no-call systems. We use the average waiting time of passengers as the main performance measure, which is defined as the average time interval between a passenger's arrival and being picked up. By Little's Law, the average waiting time of passengers is proportional to the average number of passengers waiting in the system. Thus, the average waiting time can measure passengers' satisfaction and the congestion of the system. Later, we will also consider the effective utilization level of the system (the proportion of time the taxis are driving with passengers) in Section 6.1 when passengers may abandon the services. Note that in a system without abandonment, all the above mechanisms would have the same effective utilization levels.

3.2 Observations from Numerical Experiments

In this section, we perform numerical experiments on the call and no-call systems, which will lead to some key observations. Then, we provide intuitive explanations for those

observations. In Section 3.3, we verify the observations with theoretical analysis using a novel approximation approach.

3.2.1 Observations for One-Direction Systems

We start with the one-direction call system. Throughout our study, we set $d = 10$ and $v = 1$ in the numerical experiments unless otherwise specified. Hence the service rate is fixed to be $\mu = 0.1$. Then we adjust the arrival rate λ to see how the average waiting time of passengers changes with λ. Note that this effectively changes the utilization level ρ. Furthermore, we vary the number of taxis k, the road length R, and the travel distance d to see how these parameters affect the relation between the average waiting time and the utilization level ρ. The results are shown in Figure 3.1. From Figure 3.1, we have the following observation.[2]

Observation 1 *In the one-direction call system, the average waiting time is not always monotonically increasing in ρ. In particular, it increases in ρ when ρ is small or large. However, it could decrease in ρ when ρ is medium. Moreover, such non-monotonicity is more pronounced when R is large or when d is small.*

To understand the intuition behind the non-monotonicity observed in Figure 3.1, we decompose the waiting time for one passenger into two components: the *response time* and the *en route time*. Here, the response time is defined as the time between a

[2]In all our simulation results, each average waiting time is based on the average of 500 replications of sample paths. In each sample path, we run the system from an idle state. Then we use the average waiting time between the t_0-th and t_1-th passengers as the average waiting time in that sample path. Here, t_0 can be viewed as having a warm-up period until the system enters a steady state. The value of t_0 is determined by a commonly used graphical method, see Welch (1983). In particular, let Y_j denote the average waiting time of passenger j over 500 sample paths. Then we take $w = 20$ and let $\bar{Y}_j = (1/w)\sum_{i=j-w+1}^{j} Y_i$ be the average waiting time between the $(j-w+1)$-th and the j-th passengers. We plot $\bar{Y}_j$ and observe visually when $\bar{Y}_j$ becomes stable. Then we choose a t_0 that is large enough so that $\bar{Y}_j$ has entered a steady state. We further choose $t_1 = t_0 + 6000$.

Also, we note that the observations in this dissertation are obtained from extensive numerical experiments, and the figures we choose to show are representative of the numerical results. As we will show in Section 3.3, these observations are generally valid in an approximation system.

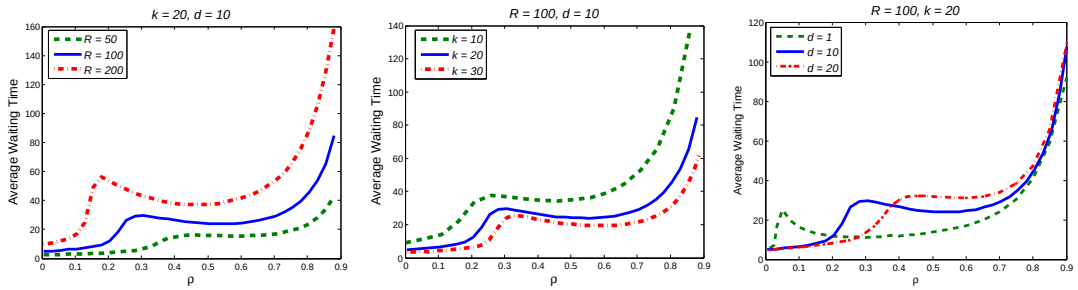


Figure 3.1: Average waiting time of the one-direction call system under different road length R, taxi number k and average ride distance d.

passenger's arrival and when the passenger's ride request is responded to by the nearest taxi, and the en route time is defined as the time between the request is responded and the passenger is picked up by a taxi. In particular, when a passenger arrives, if there are idle taxis in the system, then the passenger's ride request would be responded to immediately by the nearest taxi. In this case, the response time is zero. Otherwise, the passenger has to wait until a taxi becomes available to which the passenger is the nearest.

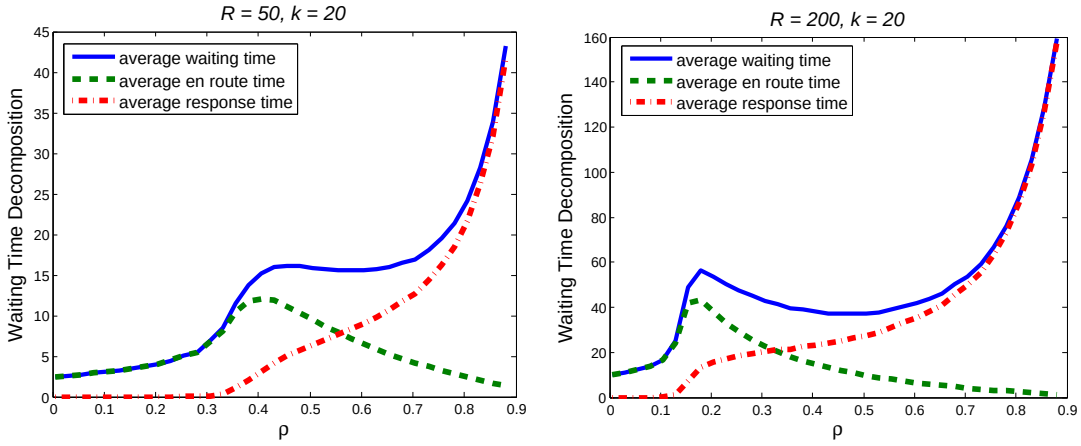


Figure 3.2: Decomposition of waiting time into response time and en route time for the one-direction call system.

Figure 3.2 shows the decomposition of the average waiting time into the average response time and the average en route time when $R = 50$, $k = 20$, and $R = 200$, $k = 20$. From Figure 3.2, we can see that the average response time is monotone in ρ while the average en route time is not. It is apparent that the non-monotonicity of the average en route time leads to the non-monotonicity of the average waiting time. To understand the non-monotonicity of the average en route time, consider the case when ρ is small. In such a case, increasing ρ leads to fewer idle taxis on average, which means that the average distance between an arriving passenger and the nearest idle taxi is longer, and thus, the en route time increases with ρ in this range. When ρ is high, a larger ρ implies that, on average, more passengers are waiting for service in the system. Therefore, the average distance between the taxi and the nearest passenger decreases in ρ; thus the en route time decreases in ρ. When ρ is in the middle range (near where the en route time peaks), a passenger arriving at the system is likely to either find him/herself among a few waiting passengers while all taxis are busy, or he/she is the only passenger waiting while there are not many taxis available in the system. In this case, the average distance between the passenger and the taxi is the largest, which explains the peak of the en route time. Moreover, this effect is more significant when R is large, because the en route time plays a more significant role in determining the total waiting time in that case. Therefore, the non-monotonicity is more pronounced when R is large. Such non-monotonicity is also more pronounced when d is small for a similar reason: the ride duration is relatively short compared to the enroute time, and thus the en route time plays a more significant role in determining the total waiting time in that case.

We next perform similar numerical experiments for the one-direction no-call system. The results are shown in Figure 3.3. From Figure 3.3, we make the following observation.

Observation 2 . *The average waiting time is monotonically increasing in ρ in the one-direction no-call system.*

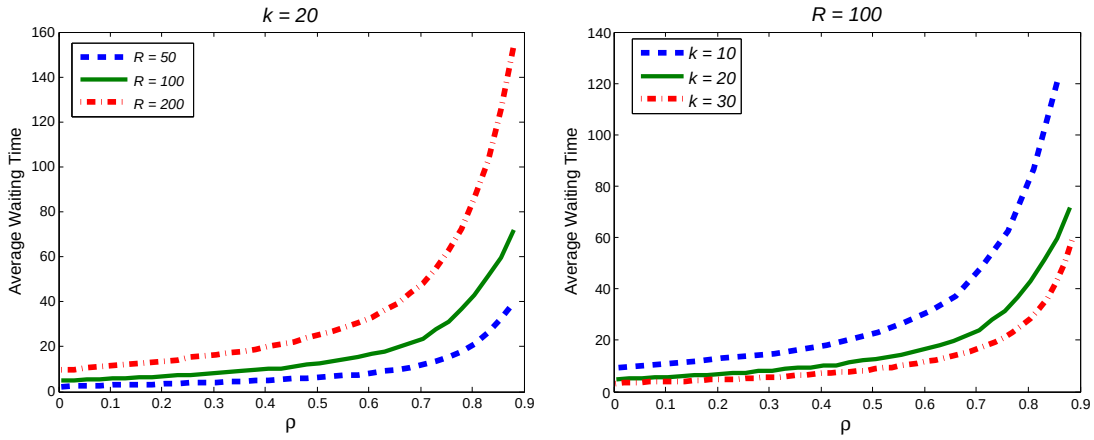


Figure 3.3: Average waiting time of the one-direction no-call system under different R and k.

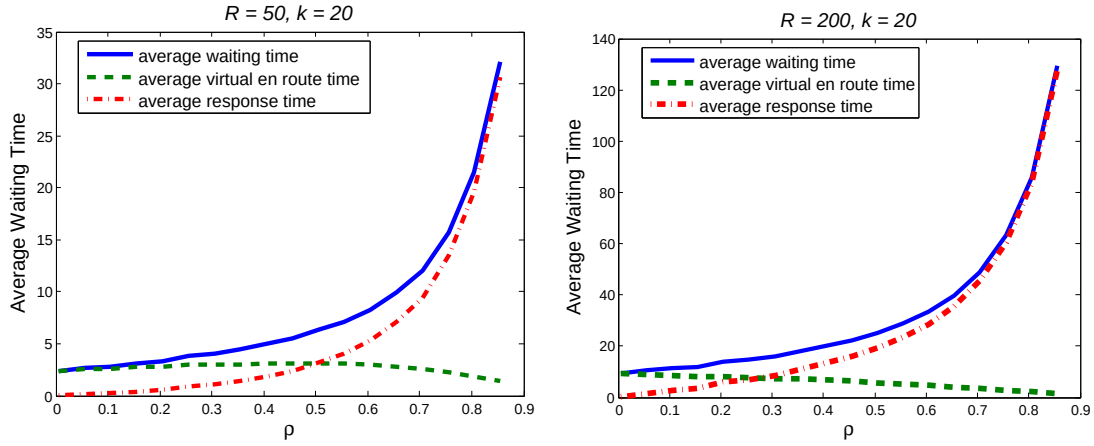


Figure 3.4: Decomposition of waiting time into response time and virtual en route time for the one-direction no-call system.

In the no-call mechanism, there are no exact counterparts of the response time and the en route time as in the call mechanism. However, similar to the en route time in the call mechanism, for each finally matched passenger-taxi pair, we define a measure named

the virtual en route time, which is the time between the passenger and the taxi becoming available and the pickup moment. The virtual en route time has similar features as the en route time in the call mechanism but also has some key differences. When ρ is small, the virtual en route time also increases in ρ since there will be fewer idle taxis as ρ increases. When ρ is large, the virtual en route time also decreases in ρ since there will be more passengers in the system as ρ increases and the average distance traveled between the time the taxi becomes available and encounters a passenger is shorter. However, the fluctuation of the virtual en route time is much smaller than that of the en route time in the call mechanism. This is because in the no-call mechanism, when the distance between a taxi and its nearest passenger is large, there is a high chance that the taxi will encounter another passenger en route and end up picking up the nearer passenger (since taxis are not committed to passengers). As a result, the virtual en route time in the no-call mechanism is much less than the en route time in the call mechanism. As we can see from Figure 3.4, the virtual en route time is much smoother than the en route time in Figure 3.2, which explains the monotonicity observed in Figure 3.3.

3.2.2 Observations for Two-Direction Systems

In this section, we consider the two-direction call/no-call systems. As in Section 3.2.1, we perform numerical experiments to study the relation between the average waiting time and the utilization level ρ under different values of k and R. We start with the call system. The results are shown in Figure 3.5. From Figure 3.5, we make the following observation.

Observation 3 *In the two-direction call system, the average waiting time is not always monotonically increasing in ρ. In particular, it increases in ρ when ρ is small or large. However, it could decrease in ρ when ρ is medium. Such non-monotonicity is more*

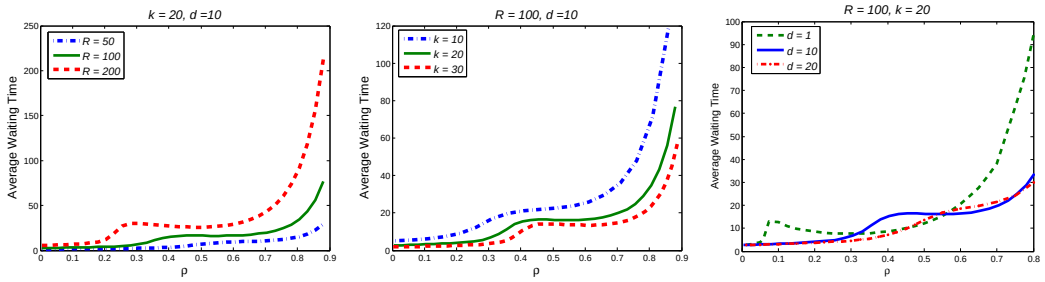


Figure 3.5: Average waiting time of the two-direction call system under different R, k and d.

pronounced when R is large or when d is small. Moreover, it is less pronounced than that in the corresponding one-direction system.

Observation 3 is similar to Observation 1 in the one-direction call system. Recall that to explain Observation 1, we decompose the average waiting time into the response time and the en route time. Here we do the same, and the result is shown in Figure 3.6. In Figure 3.6, we can see that the same phenomenon as in the one-direction system still exists in the two-direction system, which explains the non-monotonicity of the waiting time in ρ. However, as can be seen by comparing Figure 3.1 and Figure 3.5, the non-monotonicity in the two-direction call system is less significant than that in the one-direction call system. This is because in the two-direction system, the taxis can be matched to passengers in both directions; thus, the matching distance is, in general, shorter. This implies that the en route time plays a less dominant role in determining the waiting time in the two-direction systems. To further illustrate this, we compare the average en route time in one-direction versus two-direction call systems in Figure 3.7. As one can see from Figure 3.7, the non-monotonicity of the en route time is much less significant in the two-direction system.

For the two-direction no-call system, we have exactly the same result as in the one-direction system, which we summarize in the following observation.

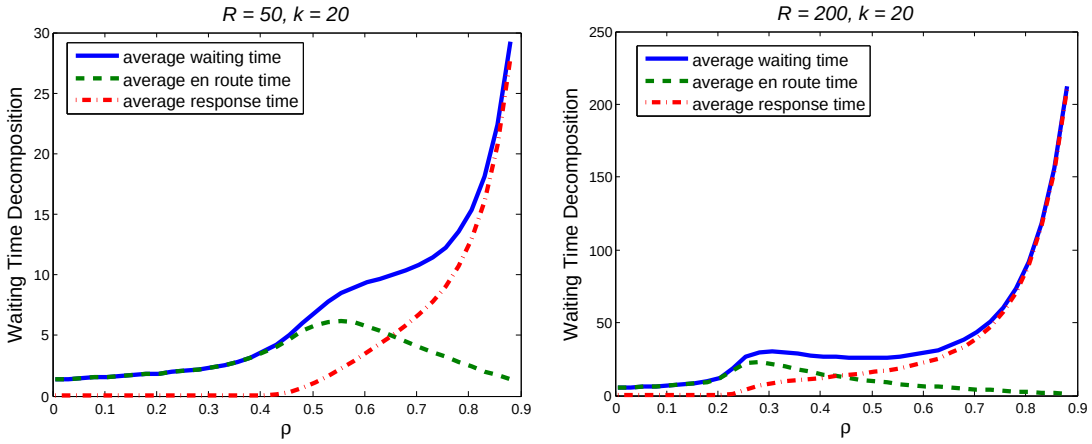


Figure 3.6: Decomposition of the average waiting time for the two-direction call system.

Observation 4 *The average waiting time in the two-direction no-call system is monotonically increasing in* ρ.

The above numerical experiments on call and no-call mechanisms lead to the key observation that the call mechanism is different from the no-call mechanism in that the average waiting time could be non-monotone in the system utilization ρ, due to the significant non-monotonicity of the en route time. In the next section, we propose an approximation scheme and provide theoretical justifications for those observations. As we shall see from the theoretical analysis, the observations we made are likely to be generally true in such systems.

3.3 Approximation Scheme

In this section, we propose an approximation scheme for the call and no-call mechanisms studied in Section 3.2. The approximation scheme is useful in several ways. First, it provides an efficient way to obtain performance measures for such systems, rather than

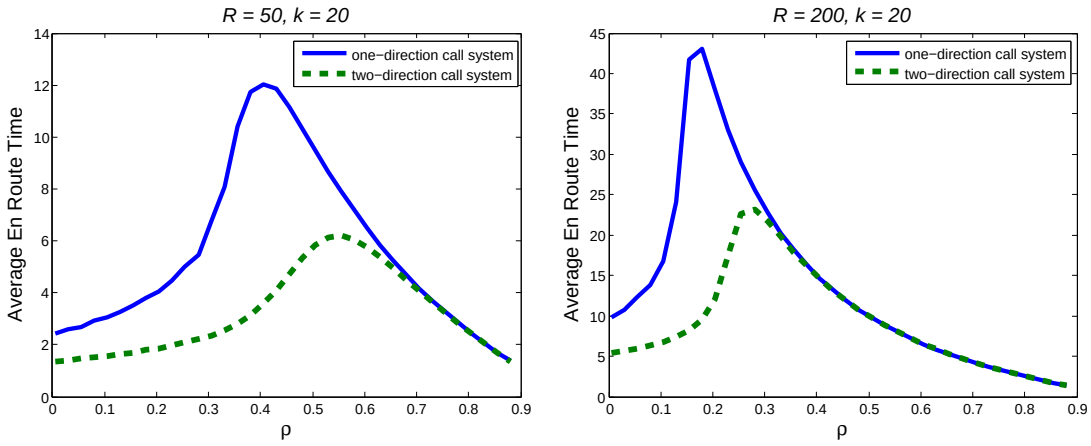


Figure 3.7: Comparison of the average en route time of the one- and two-direction call systems.

performing potentially burdensome simulations. Second, it captures the key characteristics of different matching mechanisms and thus allows us to obtain meaningful insights. As we will show, under the approximation scheme, we can prove several insights we observed in the actual system. Moreover, the approximation scheme allows extension to more complex settings, such as other road setups, passengers with limited patience, or other matching mechanisms, thus offering a powerful tool for analyzing such problems.

To establish the approximation scheme, we note that the transportation system shares some similarities with a queueing system. In particular, taxis can be viewed as servers, while passengers waiting for a ride can be viewed as customers waiting for service. The arrival rate of passengers and the service rate of taxis are also similar to the counterparts in a queueing system. However, unlike a queueing system where customers are served according to certain priority rules (e.g., first-come first-served), in the transportation system, the locations of taxis and passengers play a key role in determining the sequence of service. In addition, the waiting time for a passenger is not only impacted by waiting for busy taxis to become available, but also affected by the

en route time.

To incorporate these features, we define the *extended service time* in the transportation system, which is the sum of the actual service time and the en route time. In our model, the service time is exponentially distributed with mean $1/\mu = d/v$. However, the explicit distribution of the en route time is hard to obtain. To address this issue, in the approximation scheme, we first compute an approximate expected en route time for the next service, which we denote by $\bar{t}_e$. We note that $\bar{t}_e$ is dependent on the number of waiting passengers and available taxis in the system. Then we approximate the extended service time by assuming that it follows an exponential distribution with mean $1/\mu + \bar{t}_e$. Finally, with this approximation for the extended service time, we approximate the transportation system by an $M/M/k$ queue, in which arrivals form a single queue and are governed by a Poisson process, and there are k servers with state-dependent service rates as previously described. In the following, we specify the $M/M/k$ approximations for the one-direction and two-direction call/no-call systems.

3.3.1 Approximation Scheme for Call Mechanisms

For the one-direction call system, let n denote the total number of passengers in the system (waiting for service or being served), and $\lambda_n^{(1c)}$ and $\mu_n^{(1c)}$ denote the arrival rate and the service rate when there are n passengers in the system, respectively. By definition, $\lambda_n^{(1c)} = \lambda$ for all n. In the following, we calculate the average en route time in order to obtain $\mu_n^{(1c)}$.

In the one-direction call system, when $n < k$, the n-th passenger (the most recent arrival) would be matched to one of the $k - n + 1$ idle taxis. In the approximation scheme, we assume that the $k - n + 1$ idle taxis are uniformly located on the road, labeled by $1, 2, \ldots, k - n + 1$. Let l_i denote the clockwise distance between the i-th idle taxi and the arriving passenger. We let $t_e^{(1c)}$ represent the en route time. Then

$t_e^{(1c)} = \min_{i=1,\ldots,k-n+1} l_i/v$ is the shortest travel time between the taxis and the passenger. We have

$$\mathbb{P}(t_e^{(1c)} \geq t) = \prod_{i=1}^{k-n+1} \mathbb{P}(l_i/v \geq t) = \left(\frac{R-vt}{R}\right)^{(k-n+1)},$$

and thus, the expected en route time $\bar{t}_e^{(1c)} = \mathbb{E}[t_e^{(1c)}] = \int_0^{R/v} \mathbb{P}(t_e^{(1c)} \geq t)dt = \frac{R}{v(k-n+2)}$.

When $n > k$, there are $n-k$ passengers in the system waiting to be matched. When a taxi becomes available, it would be immediately matched to the nearest passenger. In the approximation scheme, we assume that the waiting passengers are uniformly located on the road; therefore, by a similar argument as in the $n < k$ case, we have $\bar{t}_e^{(1c)} = \frac{R}{v(n-k+1)}$. When $n = k$, we approximate the en route time by $R/2$, which is the average distance between a taxi and a passenger of random locations. To summarize, we approximate the one-direction call system by an $M/M/k$ queue with state-dependent arrival rate $\lambda_n^{(1c)} = \lambda$ and service rate

$$\mu_n^{(1c)} = \begin{cases} \frac{n}{\frac{d}{v}+\frac{R}{v(k-n+2)}}, & \text{if } n \leq k, \\ \frac{k}{\frac{d}{v}+\frac{R}{v(n-k+1)}}, & \text{if } n > k. \end{cases} \tag{3.1}$$

Next, using the approximation system given by (A.4), we prove the properties described in Observation 1. In the following, let $\mathcal{W}^{(1c)}(\lambda, d, v, k, R)$ be the average waiting time under the approximation scheme given parameters λ, d, v, k and R. We have the following result:

Theorem 3.3.1 *Consider the approximated one-direction call system given by (A.4). We have:*

1. $\mathcal{W}^{(1c)}(\lambda, d, v, k, R) < \infty$ *if and only if* $0 \leq \lambda < k\mu$, *where* $\mu = v/d$.

2. *For any given* d, v, k *and* R, *we have*

$\frac{\partial \mathcal{W}^{(1c)}(\lambda,d,v,k,R)}{\partial \lambda}\Big|_{\lambda=0} \geq 0$ *and* $\underline{\lim}_{\lambda \to k\mu-} \frac{\partial \mathcal{W}^{(1c)}(\lambda,d,v,k,R)}{\partial \lambda} > 0$.

3. *For any given* v, $k \geq 2$ *and* R, *there exists a constant* $d^*(v,k,R)$ *such that when* $d < d^*(v,k,R)$, *there exists* $0 \leq \lambda(d,v,k,R) < kv/d$ *such that*

$$\frac{\partial \mathcal{W}^{(1c)}(\lambda,d,v,k,R)}{\partial \lambda}\Big|_{\lambda=\lambda(d,v,k,R)} < 0.$$

4. *For any given* d, v *and* $k \geq 2$, *there exists a constant* $R^*(d,v,k)$ *such that when* $R > R^*(d,v,k)$, *there exists* $0 \leq \lambda(d,v,k,R) < kv/d$ *such that*

$$\frac{\partial \mathcal{W}^{(1c)}(\lambda,d,v,k,R)}{\partial \lambda}\Big|_{\lambda=\lambda(d,v,k,R)} < 0.$$

5. *If* $\mathcal{W}^{(1c)}(\lambda,d,v,k,R)$ *shows the non-monotonicity in terms of* λ *for certain travel speed* v, *then* $\mathcal{W}^{(1c)}(\lambda,d,\hat{v},k,R)$ *would also be non-monotone in terms of* λ, *where* $\hat{v}$ *denotes any arbitrary positive travel speed. Otherwise, if* $\mathcal{W}^{(1c)}(\lambda,d,v,k,R)$ *is monotonically increasing in* λ, *then* $\mathcal{W}^{(1c)}(\lambda,d,\hat{v},k,R)$ *would also be monotonically increasing in* λ.

Following a similar approach, we can approximate the two-direction call system by an $M/M/k$ queue with arrival rate $\lambda_n^{(2c)} = \lambda$ and state-dependent service rate

$$\mu_n^{(2c)} = \begin{cases} \frac{n}{\frac{d}{v}+\frac{R}{2v(k-n+2)}}, & \text{if } n \leq k, \\ \frac{k}{\frac{d}{v}+\frac{R}{2v(n-k+1)}}, & \text{if } n > k. \end{cases} \tag{3.2}$$

Using the approximation system given by (3.2), we prove the same properties for the two-direction call system.

Theorem 3.3.2 *Consider the approximated two-direction call system given by (3.2).*

Let $\mathcal{W}^{(2c)}(\lambda, d, v, k, R)$ be the average waiting time given parameters λ, d, v, k and R. Then we have:

1. *$\mathcal{W}^{(2c)}(\lambda, d, v, k, R) < \infty$ if and only if $0 \leq \lambda < k\mu$, where $\mu = v/d$.*

2. *For any given d, v, k and R, we have $\frac{\partial \mathcal{W}^{(2c)}(\lambda,d,v,k,R)}{\partial \lambda}\Big|_{\lambda=0} \geq 0$ and $\underline{\lim}_{\lambda \to k\mu-} \frac{\partial \mathcal{W}^{(2c)}(\lambda,d,v,k,R)}{\partial \lambda} > 0$.*

3. *For any given v, $k \geq 2$ and R, there exists a constant $d^*(v, k, R)$ such that when $d < d^*(v, k, R)$, there exists $0 \leq \lambda(d, v, k, R) < kv/d$ such that*

$$\frac{\partial \mathcal{W}^{(2c)}(\lambda, d, v, k, R)}{\partial \lambda}\Big|_{\lambda=\lambda(d,v,k,R)} < 0.$$

4. *For any given d, v and $k \geq 2$, there exists a constant $R^*(d, v, k)$ such that when $R > R^*(d, v, k)$, there exists $0 \leq \lambda(d, v, k, R) < kv/d$ such that*

$$\frac{\partial \mathcal{W}^{(2c)}(\lambda, d, v, k, R)}{\partial \lambda}\Big|_{\lambda=\lambda(d,v,k,R)} < 0.$$

5. *If $\mathcal{W}^{(2c)}(\lambda, d, v, k, R)$ shows the non-monotonicity in terms of λ for certain travel speed v, then $\mathcal{W}^{(2c)}(\lambda, d, \hat{v}, k, R)$ would also be non-monotone in terms of λ, where $\hat{v}$ denotes any arbitrary positive travel speed. Otherwise, if $\mathcal{W}^{(2c)}(\lambda, d, v, k, R)$ is monotonically increasing in λ, then $\mathcal{W}^{(2c)}(\lambda, d, \hat{v}, k, R)$ would also be monotonically increasing in λ.*

The proofs of all theorems are relegated to Appendix. Interestingly, the stability condition for both the one-direction and two-direction system are $\lambda < k\mu$. This is because when λ approaches $k\mu$, the en route time becomes ignorable as the distance between an available taxi and the nearest waiting passenger becomes small. In addition,

Theorem 3.3.1 and Theorem 3.3.2 provide support for Observation 1 and Observation 3 on the non-monotonicity property, respectively. In particular, the second parts of Theorems 3.3.1 and 3.3.2 show that when λ is very small (close to 0) or close to the high-traffic regime (when λ is close to $k\mu$), the average waiting time is increasing in λ. The third part of Theorems 3.3.1 and 3.3.2 state that given any R, v and k, as long as the travel distance d is small enough, there will be a range of λ such that the average waiting time is decreasing in λ. In contrast, the fourth parts of Theorems 3.3.1 and 3.3.2 state that given any d, v and k, as long as the road length R is large enough, there will be a range of λ such that the average waiting time is decreasing in λ. The last parts of Theorems 3.3.1 and 3.3.2 show that if the average waiting time is monotone (non-monotone) in the arrival rate under travel speed v, then the average waiting time is also monotone (non-monotone) in the arrival rate under any other arbitrary travel speed $\hat{v}$.

Therefore, with the help of the approximation scheme, we are able to verify that the findings we obtained in the simulation are generally true in such systems. These theorems further verify that the non-monotonicity of the average waiting time is closely related to the large ratio of the enroute time and the ride duration.

3.3.2 Approximation Scheme for No-Call Mechanisms

Next, we propose an approximation scheme for the no-call mechanisms. Since the matching mechanism for the one-direction no-call system is the same as that for the two-direction no-call system except that taxis are allowed to travel in both directions, the adopted approximation schemes for the two systems are identical. In the following, we apply the same idea as in the call mechanism and approximate the system with a state-dependent $M/M/k$ queue. Let n denote the number of passengers in the system and $\lambda_n^{(nc)}$ and $\mu_n^{(nc)}$ denote the arrival and service rates when there are n passengers in

the system. Apparently, the arrival rate $\lambda_n^{(nc)} = \lambda$. We now approximate the average en route time $\bar{t}_e^{(nc)}$ when there are n passengers in the system, in order to obtain the service rate $\mu_n^{(nc)}$.

When $n > k$, there are more passengers in the system than the total number of taxis. Though the number of taxis in service could be smaller than k, we simplify the analysis by assuming that all k taxis are in service and $n - k$ passengers are waiting for service. Let $t_e^{(nc)}$ be the en route time for the next available taxi in this case. The event $t_e^{(nc)} \geq t$ implies that no passenger is located within clockwise distance vt of the taxi, and the taxi does not encounter any new passenger arriving within time period t. Here we ignore the possibility of another taxi becoming available during this interval and picking up a passenger before this taxi. Let l_i denote the distance between the i-th waiting passenger and the taxi. Based on the discussions above, we have

$$\mathbb{P}(t_e^{(nc)} \geq t) = \mathbb{P}\left(\min_{1 \leq i \leq n-k} l_i/v \geq t\right) \mathbb{P}(A),$$

where A denotes the event that the taxi encounters no new passenger arriving within time period t (note that the event A depends only on future arrivals and thus is independent of the event $\min_{1 \leq i \leq n-k} l_i/v \geq t$).

Similar to the analysis for the call mechanism, it is easy to see that $\mathbb{P}(\min_{1 \leq i \leq n-k} l_i/v \geq t) = (\frac{R-vt}{R})^{n-k}$. Now we compute the probability of event A. We note that the new passenger arrivals within time period t follow a Poisson process with arrival rate λ. For an arrival at time $\tau \leq t$, to be encountered by the taxi before time t, the passenger must be within a travel distance of $v(t - \tau)$. Since the arrival location is uniformly distributed, the probability that an arrival is in a region of length $v(t - \tau)$ is $\frac{v(t-\tau)}{R}$; therefore, the number of arrivals that would be encountered by the taxi within time t follows a Poisson process with non-homogenous arrival rate $\lambda_\tau = \lambda \frac{v(t-\tau)}{R}$ at any time

$\tau \leq t$. Consequently, the total number of arrivals encountered by the taxi within time t, which we denote by N_t, follows a Poisson distribution with mean $\int_0^t \lambda \frac{v(t-\tau)}{R} d\tau = \frac{\lambda v t^2}{2R}$. Therefore,

$$\mathbb{P}(A) = \mathbb{P}(N = 0) = \exp\left(-\frac{\lambda v t^2}{2R}\right),$$

and thus, the expected en route time

$$\bar{t}_e^{(nc)} = \mathbb{E}[t_e^{(nc)}] = \int_0^{R/v} \mathbb{P}\left(t_e^{(nc)} \geq t\right) dt = \int_0^{R/v} \left(\frac{R - vt}{R}\right)^{n-k} \exp\left(-\frac{\lambda v t^2}{2R}\right) dt.$$

Now we analyze the case when $n \leq k$. We assume that $n-1$ taxis are already in service, the n-th passenger would be picked up by one of the $k - n + 1$ taxis that are idle. Now we approximate the waiting time for a new arriving passenger, which is also the en route time $t_e^{(nc)}$ for a taxi picking this passenger up. If we further assume no other passenger arrives before this passenger is picked up, then the event $t_e^{(nc)} \geq t$ is equivalent to that no idle taxi is within vt clockwise distance from the passenger and no busy taxi frees up within $v(t-\tau)$ clockwise distance at any time point $\tau \leq t$. Let B denote the latter event (no busy taxi frees up within $v(t-\tau)$ distance at any time point τ satisfying $\tau \leq t$), and let l_i denote the distance between the i-th idle taxi and the passenger. By similar arguments as in the $n > k$ case, we have

$$\mathbb{P}(t_e^{(nc)} \geq t) = \mathbb{P}\left(\min_{1 \leq i \leq k-n+1} l_i/v \geq t\right) \mathbb{P}(B) = \left(\frac{R - vt}{R}\right)^{k-n+1} \exp\left(-\frac{(n-1)v^2t^2}{2Rd}\right),$$

and $\bar{t}_e^{(nc)} = \mathbb{E}[t_e^{(nc)}] = \int_0^{R/v} \mathbb{P}(t_e^{(nc)} \geq t) dt = \int_0^{R/v} (\frac{R-vt}{R})^{k-n+1} \exp(-\frac{(n-1)v^2t^2}{2Rd}) dt$.

Thus, the no-call mechanism can be approximated by an $M/M/k$ queue with arrival

rate $\lambda_n^{(nc)} = \lambda$ and state-dependent service rate

$$\mu_n^{(nc)} = \begin{cases} \frac{n}{d/v + \int_0^{R/v} \left(\frac{R-vt}{R}\right)^{k-n+1} \exp\left(-\frac{(n-1)v^2t^2}{2Rd}\right) dt}, & n \leq k, \\ \frac{k}{d/v + \int_0^{R/v} \left(\frac{R-vt}{R}\right)^{n-k} \exp\left(-\frac{\lambda v t^2}{2R}\right) dt}, & n > k. \end{cases} \tag{3.3}$$

In the following, let $\mathcal{W}^{(nc)}(\lambda, d, v, k, R)$ be the average waiting time under the approximated no-call mechanism. We have the following theorem:

Theorem 3.3.3 *For any d, v and k, there exists $R^*(d, v, k)$ such that when $R < R^*(d, v, k)$, $\mathcal{W}^{(nc)}(\lambda, d, v, k, R)$ is increasing in λ.*

Theorem 3.3.3 provides support for Observations 2 and 4 for the no-call mechanism under the approximation schemes. By Theorem 3.3.3, in the approximated no-call mechanism with a short road length, the waiting time is always increasing in λ, which is consistent with both Observations 2 and 4.[3] Thus, by using the approximation scheme, we are able to provide justifications for most of the observations about the no-call mechanisms.

3.3.3 Evaluation of Approximations

In this section, we evaluate the performance of our approximation schemes by computing the average waiting time given by the approximation scheme and comparing it with the average waiting time of the true system. The results are shown in Figures 3.8, 3.9 and 3.10. As we can see, the approximation scheme provides a good approximation for the average waiting time of the true system, except for large values of ρ. In particular, it retains the important features of the original system. We note that in most cases, the approximation system gives an underestimate of the average waiting time compared to the true system, especially when the utilization is high. This may be because we

[3]Unfortunately, we are not able to prove this statement for an arbitrary value of R.

"reduced" the variance in the system by assuming the idling taxis and the waiting passengers are uniformly located on the road and the extended service time follows an exponential distribution (the true extended service time tends to have a longer tail). Such reduction in variance in our modeling leads to the smaller estimation of the average waiting time when the utilization is high. Such observation is consistent with the finding in Gupta et al. (2010), who show that it is generally true that the approximation of M/G/K queueing system may not perform well when the system utilization or the variance of the service time is high.

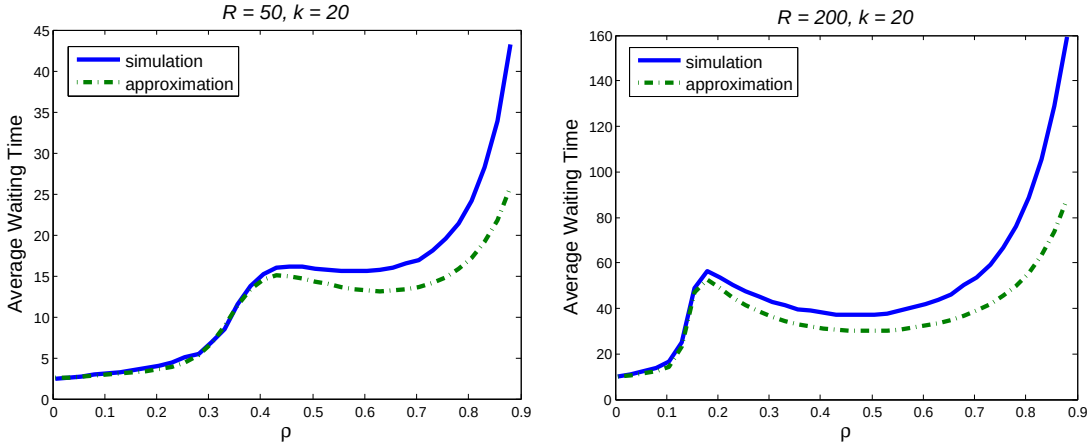


Figure 3.8: Performance of the approximation scheme for the one-direction call system.

We further compare our approximation with an approximation scheme proposed by McLeod (1972). We show that our approximation for the first time captures the non-monotonicity feature of the call mechanism. In McLeod (1972), the author considers a "many-to many taxi operations" approximation scheme. In the following, we first show how that scheme can be adapted to the circular road network considered in our work. In McLeod (1972)'s approximation, it assumes that each cab operates independently of the other cabs and serves its share of incoming requests. To adapt it to our case, the circular road R is evenly divided into k regions and the requests in each effective service region

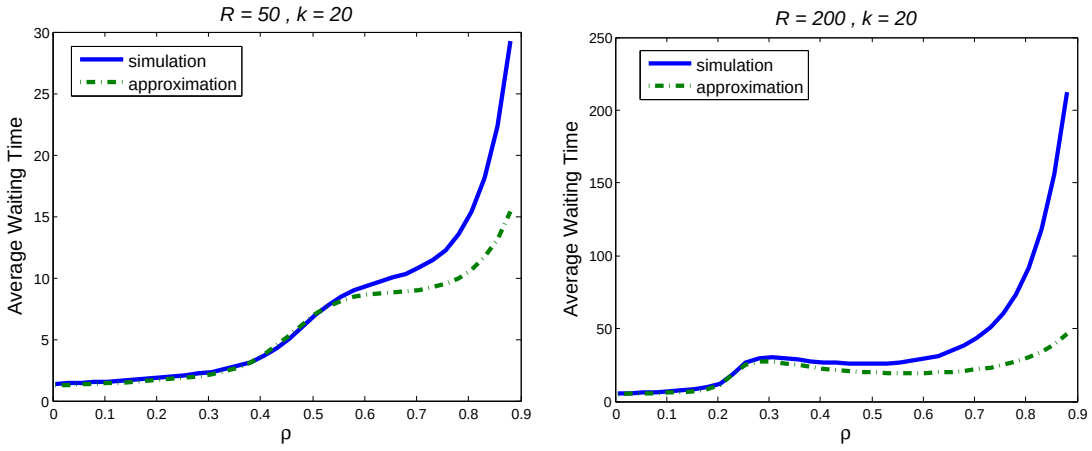


Figure 3.9: Performance of the approximation scheme for the two-direction call systems.

R/k is responded by only one taxi dedicated to that region. When a taxi becomes free, it may be at any location in its effective service area, and the next passenger awaiting pickup will also be randomly located within the area. The average distance between the taxi and the next passenger can be calculated to be $R/3k$. The taxi service rate, which includes time for en route and with passenger in the taxi, is $1/(1/\mu + R/3k)$; the passenger arrival rate for each taxi is λ/k. Therefore, an $M/M/1$ queueing model with arrival rate λ/k and service rate $1/(1/\mu + R/3k)$ is used to approximate this system. And the passenger's average waiting time is given by $W_q + R/3k$, where W_q is the waiting time of the $M/M/1$ model.

Figure 3.11 compares our approximation scheme and the one proposed in McLeod (1972), where $d = 10$, $v = 1$ and $\mu = v/d = 0.1$. It illustrates that the average waiting time in the approximation scheme proposed in McLeod (1972) monotonically increases with the system utilization level, while our model provides a much more accurate approximation by capturing the non-monotonicity property.

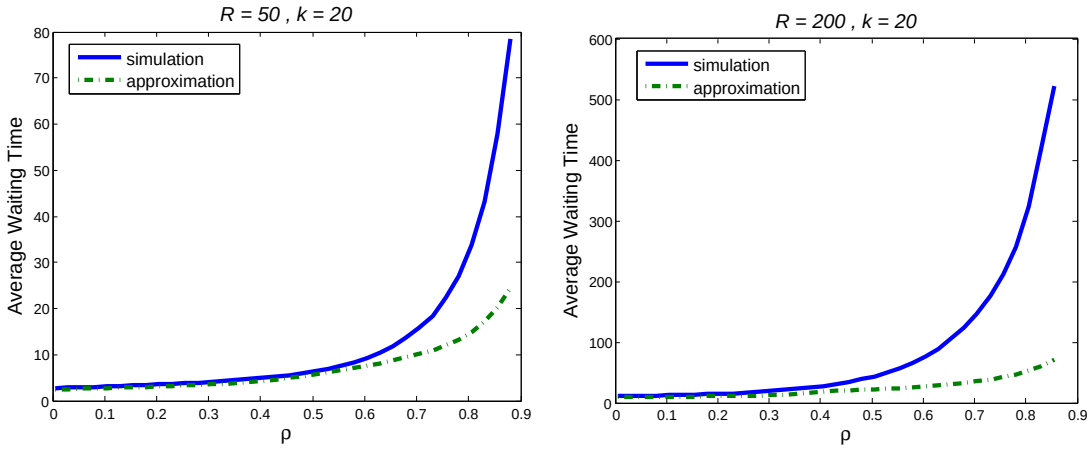


Figure 3.10: Performance of the approximation scheme for the two-direction no-call systems.

3.4 Comparison Between Call and No-Call Mechanisms

In this section, we compare the average waiting time between call and no-call mechanisms in both one and two-direction systems.

We first compare the one-direction call and no-call systems and the result is shown in Figure 3.12(a). From Figure 3.12(a), we observe that the average waiting time in the one-direction call system is always larger than that in the corresponding no-call system. Indeed, using the approximation scheme introduced in Section 3.3, we are able to prove the following theorem that verifies this observation.

Theorem 3.4.1 *For any* λ, d, v, k *and* R, $\mathcal{W}^{(nc)}(\lambda, d, v, k, R) \leq \mathcal{W}^{(1c)}(\lambda, d, v, k, R)$.

To explain this result, we note that in the one-direction case, the no-call mechanism always expedites the next pickup compared to the call mechanism. More specifically, in the call mechanism, the nearest passenger and taxi are matched as soon as available; in contrast, in the no-call mechanism, a passenger-taxi match is not established until the taxi passes by a waiting passenger, which provides the possibility of more efficient

matches. In particular, the advantage of the no-call mechanism can be demonstrated with two possible scenarios. For the current nearest pair of passenger and taxi: 1) the passenger might be picked up by another taxi that becomes available later and has a shorter distance to the passenger upon its availability; 2) the taxi may encounter a new arriving passenger with a shorter distance to the taxi. In either scenario, the no-call mechanism shortens the next pickup time by postponing the decision-making time and thus achieves shorter average waiting time.

In addition, the difference between the call and no-call mechanisms is most significant around the point where the call mechanism has the highest average en route time. We note that when ρ approaches zero, the average waiting times for both systems are the same. This is because when ρ approaches zero, all k taxis are available when a passenger arrives, and few new passengers would arrive before the next pickup. Therefore, call and no-call mechanisms would provide the same match and thus have the same waiting time. We also note that when ρ approaches one, the average waiting times in the two systems converge and go to infinity. This is because when ρ is large, the number of waiting passengers in both systems is large, and thus, the en route time is negligible compared to the response time. Therefore, the difference between the call and no-call mechanisms is also small.

Next, we compare the two-direction call and no-call systems. The results are shown in Figure 3.12(b). As shown in Figure 3.12(b), the average waiting time in the two-direction call system is larger than that in the two-direction no-call system when ρ is medium, but is smaller than that in the two-direction no-call system when ρ is either small or large. Again, using the approximation scheme introduced in Section 3.3, we are able to verify part of this observation.

Theorem 3.4.2 *When comparing the approximation systems for two direction call and no-call systems, we have the following results:*

1. For any d, v, k and R, we have that

$$\lim_{\lambda \to 0^+} \mathcal{W}^{(nc)}(\lambda, d, v, k, R) > \lim_{\lambda \to 0^+} \mathcal{W}^{(2c)}(\lambda, d, v, k, R).$$

2. For any d, v, k and R, we have that

$$\lim_{\lambda \to kv/d^-} \mathcal{W}^{(nc)}(\lambda, d, v, k, R) > \lim_{\lambda \to kv/d^-} \mathcal{W}^{(2c)}(\lambda, d, v, k, R).$$

To understand this result, we pinpoint the disadvantage of both mechanisms and study when these disadvantages are most significant. For the call mechanism, its disadvantage can be illustrated by the following scenario: Suppose there is one waiting passenger and one idle taxi in the system, and their distance is large. According to the call mechanism, the passenger will be immediately matched to the taxi. However, such a matching is unlikely to be the most efficient one since the taxi may well encounter a new arrival passenger while it is en route to the matched passenger, or a taxi closer to the passenger may become available later. In either case, the system would have been more efficient had the matching not happened. In particular, such an undesirable scenario is more likely to occur when the road perimeter is large (thus, the en route distance is relatively long compared to the ride distance) and when the numbers of unserved passengers and idle taxis are small in the system. The latter occurs when the utilization ρ is in the middle range. We further note that a no-call mechanism would not have such inefficiency since the matching would not have been established and the taxi (passenger, respectively) would be free to pick up new passengers (be picked up by other taxis, respectively). However, the no-call mechanism suffers the disadvantage that a taxi does not know the shortest distance to drive to pick up the nearest passenger, and thus may have a chance to miss the passenger by driving to the opposite direction.

This disadvantage of the no-call mechanism dominates the advantage when ρ is either small or large, resulting in a longer average waiting time in the no-call mechanism in these cases.

In this chapter, we have investigated the characteristics of the call and no-call mechanisms and pinpointed their advantages and disadvantages. In the next Chapter, we propose an improved matching mechanism by combining the advantages of both mechanisms.

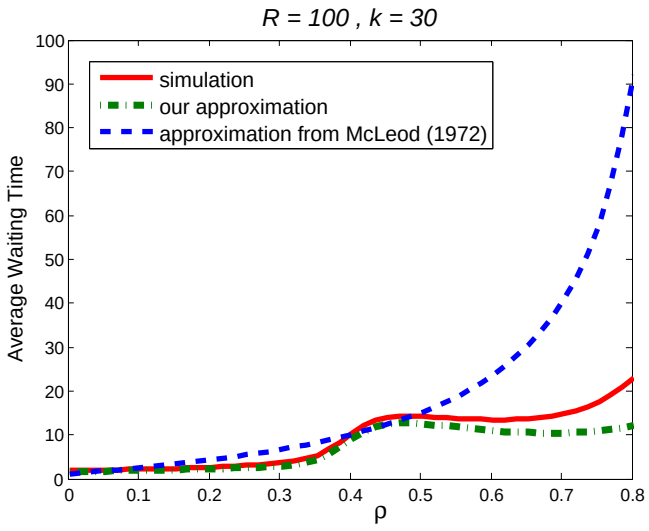
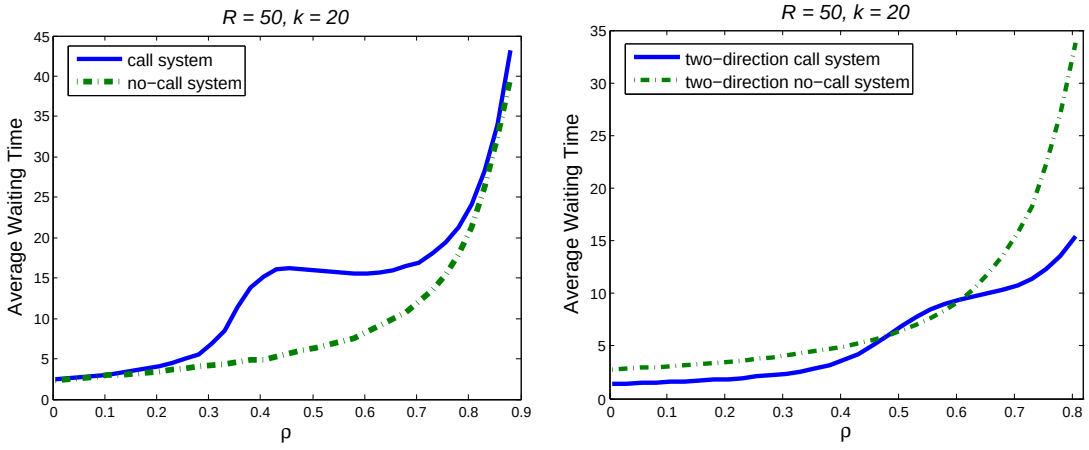


Figure 3.11: Comparison between our approximation of the two direction call system with that in McLeod (1972).



(a) One-direction system

(b) Two-direction system

Figure 3.12: Comparison of average waiting time between call and no-call systems.

Chapter 4

An Improved Matching Mechanism

In this chapter, we focus on the two-direction systems as they are closer to reality. Recall that in the two-direction call system, the taxis have information about which driving direction is more efficient, but by matching to a passenger immediately, there is a possibility of missing a later-arrival passenger with a shorter pickup distance. This can be avoided in the two-direction no-call system. However, taxis in the two-direction no-call system lack the information about the location of the nearest passenger and thus could miss a nearby passenger by driving in the other direction.

In this chapter, we propose a modified mechanism that exploits the advantages and mitigates the disadvantages of the call and no-call mechanisms. In particular, we consider a call mechanism with a response cap. Specifically, the mechanism runs like the call mechanism except that a taxi and a passenger will not be matched unless their distance is within a designated cap. We call this mechanism the *capped matching mechanism* and the corresponding transportation system the *capped system*. The focus of this chapter is on studying the effect of adding such a cap on the average waiting

time, as well as proposing a heuristic method to calculate a near-optimal cap. In theory, the cap could be chosen according to the state of the system. In the extreme case, the cap can be chosen such that the decision of whether to match a passenger and a taxi can depend on the entire history of the system. Such more complicated controls could improve the performance of the system. However, they will be very complicated to calculate or implement. Therefore, in this chapter, we focus on the case where the cap is a fixed constant. We will see that even such a relatively simple added lever could improve the efficiency of the system significantly.

To start, we perform numerical tests of the capped matching mechanism under different caps c (from 0 to $R/2$) and find the cap that leads to the smallest average waiting time. Note that $c = 0$ corresponds to the no-call mechanism and $c = R/2$ (the maximum distance between a taxi and a passenger in a circular road) corresponds to the call mechanism. The performance of the capped system with the optimal cap is shown in Figure 4.1.

From Figure 4.1, we can see that the capped system with the optimal cap outperforms the call and no-call systems under different parameter settings. In particular, using an optimal cap could significantly reduce the average waiting time, and the reduction is more significant when there are more taxis in the system.

Having observed the potential value of adding a cap, a natural question is how to obtain a good cap without having to go through extensive simulations. In the following, we propose a heuristic method that provides a cap with good performance.

To find a good cap such that we should match a taxi and a passenger if their distance is within the cap and should not otherwise, we consider the following question: How close do a taxi and a passenger need to be so that matching them would be a *good* matching? To answer this question, we think from the taxi's point of view. Recall that if we establish a match between a taxi and a passenger, then the en route time

for this taxi will be equal to the distance between the taxi and the passenger. Now we approximate the expected time this taxi has to drive in order to encounter a passenger if we do not establish the match. If the expected time is shorter than the distance from the current passenger, then intuitively the current match is not efficient, in which case we should not match them. Otherwise, matching will result in a shorter en route time than not matching, in which case we should match them.

Now it remains to calculate the expected time the taxi has to drive in order to encounter a passenger. To this end, we first compute the probability that a new passenger arrival will be encountered by this taxi within time x. Remember that new passengers arrive according to a Poisson process with rate λ. For an arrival at time y, $y \leq x$, to be encountered by the taxi before time x, the passenger must be located between y and x, which is of length $x - y$. By the assumption of the uniformly distributed arriving location, the probability that an arrival will be in a region of length $x - y$ is $\frac{x-y}{R}$; therefore, the arrivals that would be encountered by the taxi within time x follow a Poisson process with a non-homogenous arrival rate $\lambda_y = \lambda\frac{x-y}{R}$ at any time $y \leq x$. Consequently, the total number of arrivals encountered by the taxi within time x, which we denote by N, follows a Poisson distribution with mean $\int_0^x \lambda\frac{x-y}{R}dy = \frac{\lambda x^2}{2R}$. Therefore, the probability that it takes longer than x for a taxi to encounter a passenger equals

$$\mathbb{P}(N = 0) = \exp\left(-\frac{\lambda x^2}{2R}\right),$$

and the expected time for a taxi to encounter a passenger is

$$\int_0^\infty \exp\left(-\frac{\lambda x^2}{2R}\right) dx = \sqrt{\frac{\pi R}{2\lambda}}.$$

Thus, based on the above idea, we choose a heuristic cap c^* to be $c^* = \min\left\{R/2, \sqrt{\frac{\pi R}{2\lambda}}\right\}$.

Here, we add the term $R/2$ because for any $c^* \geq R/2$, the mechanism will be equivalent to a call mechanism.

Next, we test the performance of our proposed heuristic cap. We define the extra average waiting time as the waiting time difference divided by the waiting time in the optimal capped system. Figure 4.2 shows the extra average waiting time (in percentage) of the heuristic capped system, call and no-call systems compared to the optimal capped system.

From Figure 4.2, we can see that the average waiting time in the no-call/call system could be triple the average waiting time in the optimal capped system in the worst case. However, the heuristic cap performs quite well, with no more than 12% extra average waiting time versus using the optimal cap. Therefore the heuristic capped system has much better performance than both the call and no-call systems. And such a mechanism is also very easy to implement in practice.

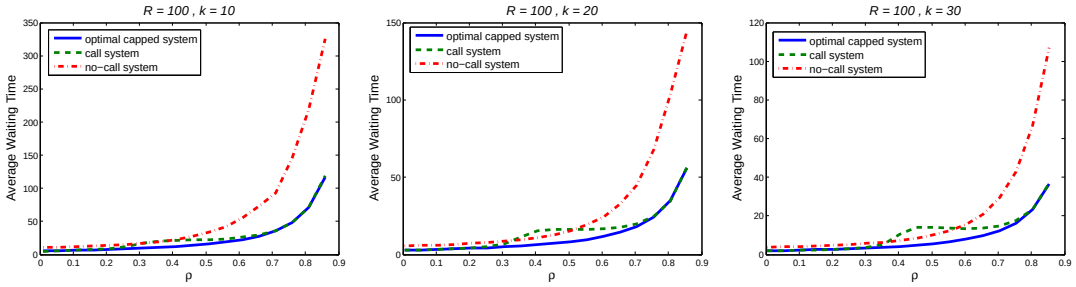


Figure 4.1: Comparison of the two-direction optimal capped system, no-call system and call system.

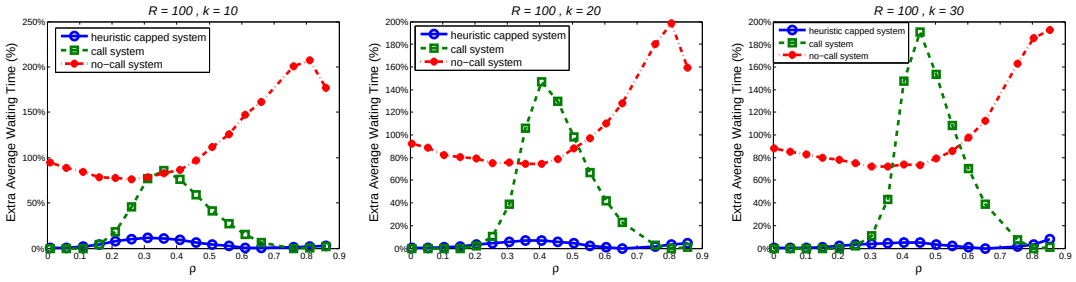


Figure 4.2: Extra average waiting time (%) of call and no-call systems compared to the optimal capped system.

Chapter 5

More Complicated Road Network

In this chapter, we extend our numerical study to a more realistic setting: a two-dimensional grid network. The two-dimensional grid network provides a closer model for city road networks and has been used extensively in transportation-related studies (see, e.g., Fawaz and Newell (1976)). In particular, we consider a rectangular region that is divided into equally sized sub-squares by vertical and horizontal roads. The sub-squares may be viewed as blocks in the city. Mathematically, an $m \times n$ two-dimensional grid network is comprised of roads of $\{x = k, 1 \leq y \leq n\}$, $k = 1, \ldots, m$ and $\{y = l, 1 \leq x \leq m\}$, $l = 1, \ldots, n$. In the numerical experiments, we simulate the call and no-call systems in the two-dimensional grid network to see if the characteristics of these systems in the circular road setting still hold.

In the simulation, we assume that the passengers' arrival still follows a Poisson process and they arrive only at intersection points of the road, i.e., passengers arrive only at points (x, y) where $1 \leq x \leq m$ and $1 \leq y \leq n$ are both integers. This is mainly for the simplicity of analysis, yet it is also plausible in practice. Taxis drive on the road, and at each intersection point, they choose either to keep the current direction or turn left or right (each option is available only when it is eligible) in a uniformly random

manner. Note that under this setup, a taxi would not drive back and forth between two adjacent intersection points, which is reasonable and intuitively more efficient in practice. We assume taxis drive one block per unit of time, and the system is updated at each integer point of time. In the no-call system, at the end of each time epoch, if an available taxi encounters a passenger at an intersection, then a match between this taxi and the passenger will be established immediately. In the call system, at the end of each time epoch, if there are available taxis in the system, then the passengers who arrive within the past unit of time will be matched one by one to the taxis in a first-come-first-serve manner. In either system, regardless of the pickup point, the passenger's destination is uniformly distributed over all intersection points in the grid. For the service rate, we first calculate the average distance (1-norm distance) between any two uniformly random points on the grid and choose μ to be the reciprocal of the average distance. By simple calculation, we have $\mu = 3mn/(m(n^2-1)+n(m^2-1))$.

In the following, we perform numerical experiments using grid networks of different sizes and with different numbers of taxis. As shown in Figure 5.1, the main features that we have observed in the circular road setting still hold. In particular, the no-call mechanism could be more efficient than the call mechanism when the utilization is in the middle range. Otherwise, the call mechanism is more efficient. Moreover, we observe that in the grid network, the call mechanism tends to be more efficient when there are few taxis in the system or when the grid is large. This is because if the grid is large (with many roads in the grid), then the call mechanism would be more advantageous in helping taxis find passengers. Imagine a network with numerous roads; in that case, it is hard for a taxi to find a passenger without using a call mechanism. Similarly, when there are very few taxis, it is hard for them to find passengers without using the call mechanism too. Thus, the call mechanism is more efficient in those settings.

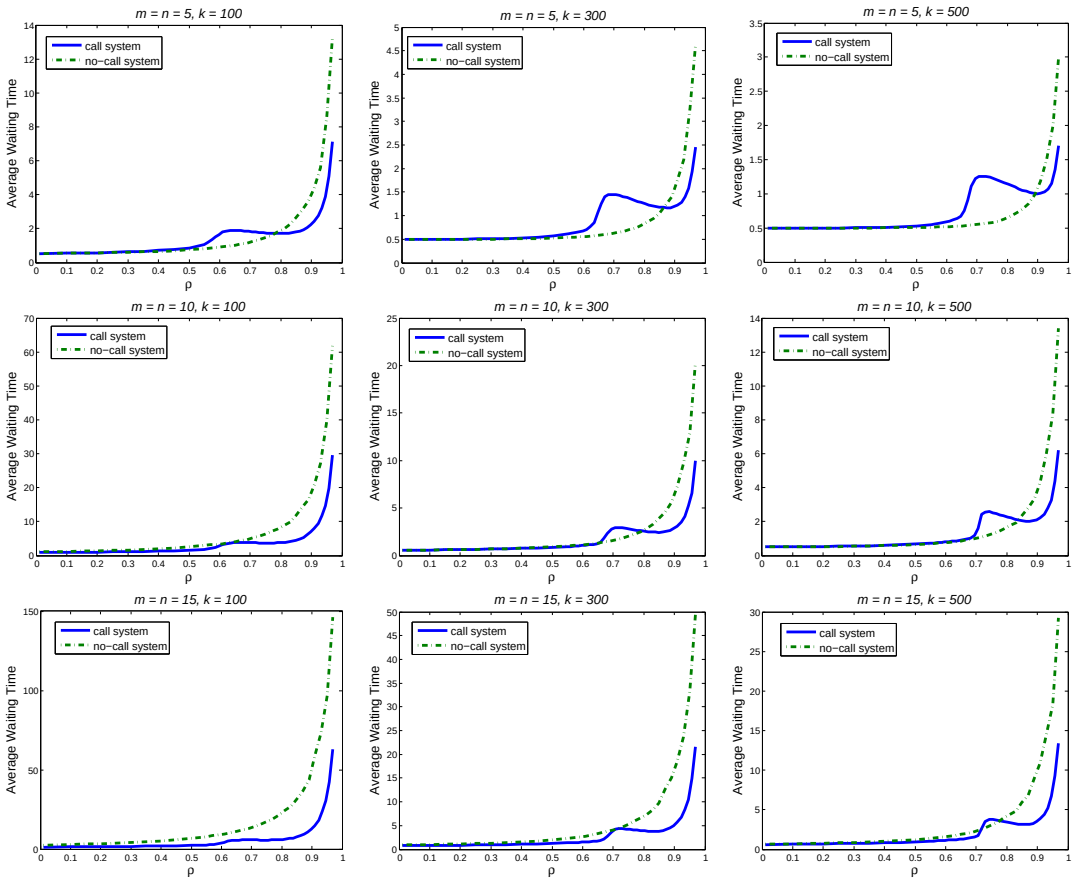


Figure 5.1: Numerical results for grid networks.

Chapter 6

Model Extensions

In this chapter, we consider some extensions of our model. We discuss how the performance of the call and no-call systems would be affected (1) when passengers have limited patience in waiting; (2) when taxis in the call mechanism could stay at stands when idle; and (3) when the travel speed of the taxis depends on the traffic congestion level of the system. These extensions not only show that the inefficiency of the call mechanism that we have observed in the basic model still exists, but also provide more insights to the operational performance of the call and no-call mechanisms under different settings.

6.1 Passenger Abandonments

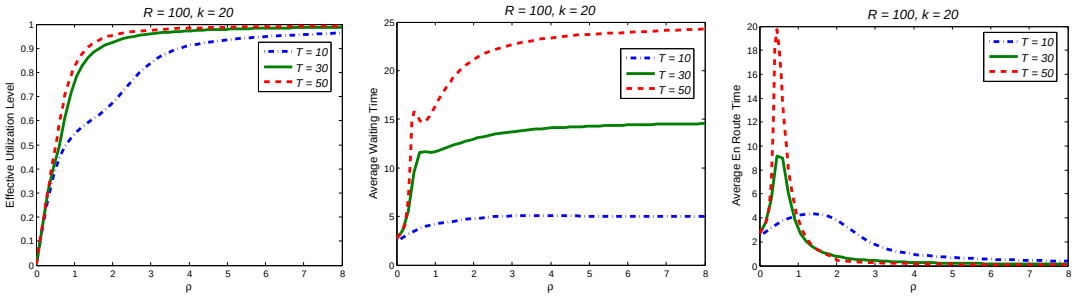
In this section, we examine the performance of two-direction call and no-call systems when passengers have limited patience in waiting. All the assumptions remain the same as in Section 4, except that each passenger is only willing to wait up to a constant patience level T, and will abandon the service once the waiting time exceeds (or is expected to exceed) T. More specifically, in the no-call mechanism, a passenger would leave the system if not picked up by time T. In the call mechanism, a passenger would

leave the system if not responded by time T. In the case that a passenger is responded at time $t < T$, the passenger will abandon the service if the en route time is longer than the remaining patience $T - t$.

We first examine the *effective utilization level* in the call mechanism under different patience levels. The effective utilization level is defined as the proportion of time that taxis spend in transporting passengers. It does not include the en route time. By queueing theory, the effective utilization level could also be represented by $\rho' = \frac{\lambda(1-\theta_a)}{k\mu}$, where θ_a is the abandonment rate of the system. Note that the system utilization $\rho = \lambda/k\mu$ can be larger than 1 when passengers may abandon the services, while the effective utilization ρ' cannot. A higher effective utilization level is more favorable to taxi drivers because it represents a higher service level of the system. In Figure 6.1(a), we show the effective utilization level in the two-direction call systems when $T = 10, 30$ and 50. As expected, the effective utilization level ρ' increases with the system utilization ρ, and it is higher when passengers are more patient since fewer passengers would abandon the services.

We next examine the average waiting time in the two-direction call system under different patience levels. The results are shown in Figures 6.1(b) and 6.1(c). In Figure 6.1(b), we observe that the non-monotonicity of average waiting time in the call system still exists with passenger abandonment, especially when the patience level is large. In addition, the average waiting time increases in the patience level T. This is intuitive as passengers can wait longer with higher patience level. Figure 6.1(c) illustrates the reason for the non-monotonicity by showing the average en route time in the call system under different patience levels T. As we can see, the non-monotonicity of the en route time still plays a key role in contributing to the non-monotonicity of the average waiting time. In particular, the non-monotonicity of the en route time is more significant when

T is large (note that when $T = \infty$, the system is equivalent to the original call system), contributing to the more significant non-monotonicity of the average waiting time. When T is small, the en route time is much smoother, which leads to less significant non-monotonicity of the average waiting time.

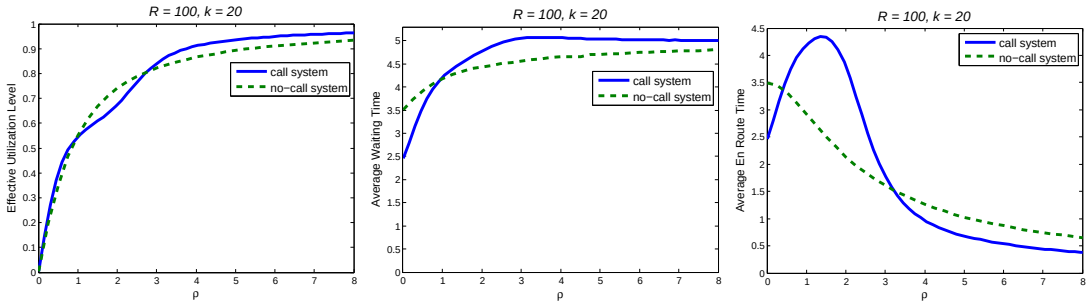


(a) Effective utilization level (b) Average waiting time (c) Average en route time

Figure 6.1: Metrics of the two-direction call system under different patience levels.

We next compare the operational performance of the two-direction call and no-call systems with passenger abandonments. Recall that when passengers have unlimited patience, the call and no-call systems have the same effective utilization level since the percentage of abandonment equals zero. It is no longer true when passengers have limited patience since the abandonment rate becomes different in these two systems. Figure 6.2 compares the performance of two-direction call and no-call systems when passengers' patience level $T = 10$. When ρ is in the middle range, the call system has a lower effective utilization level but a longer average waiting time than the no-call system. It suggests that in this case, the call mechanism is inefficient compared to the no-call mechanism from both the perspective of passengers and the perspective of drivers. Interestingly, unlike the standard queueing model with abandonments, longer average waiting time does not imply a higher effective utilization in this case. This is because the longer waiting time in the call mechanism is not driven by less abandonments, but by longer en route time. However, when the system utilization level is very low, the

call mechanism has the advantage of directing taxis to pick up the nearest passengers, resulting in slightly higher effective utilization and shorter average waiting time. When the system utilization level is high, the same advantage results in less abandonments, and thus higher effective utilization and longer waiting time.

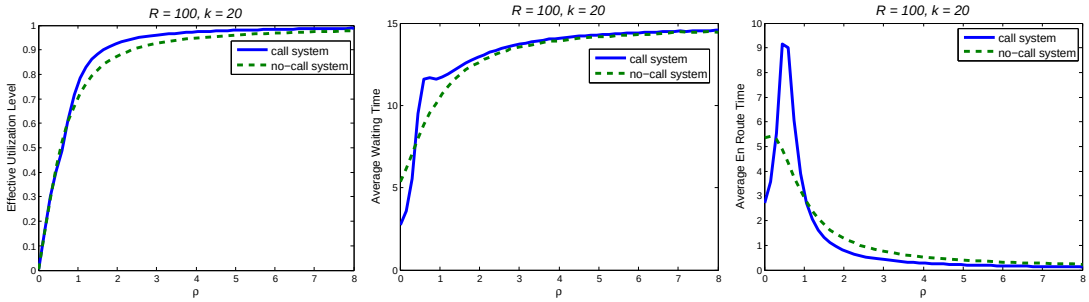


(a) Effective utilization level (b) Average waiting time (c) Average en route time

Figure 6.2: Comparison of the two-direction call and no-call systems with patience level $T = 10$.

Similar phenomenon is observed when $T = 30$ in Figure 6.3. The average waiting time in the call system is higher than the no-call system when the system utilization level is high. However, when comparing Figure 6.3(a) with Figure 6.2(a), we find that the range in which the call mechanism results in lower effective utilization level compared to the no-call mechanism shrinks. This is because the peak value of the en route time in the call system with $T = 30$ occurs at a relatively lower ρ compared to that in the call system with $T = 10$, which results in fewer passenger abandonments compared to the case when passengers are less patient.

In sum, although passenger abandonments may help reduce the peak en route time in the call mechanism, the non-monotonicity of the average waiting time still exists and our main findings in previous sections are still valid. Furthermore, with passenger abandonments, the call mechanism may result in lower effective utilization level compared to the no-call mechanism. Such an observation is consistent with the empirical evidence



(a) Effective utilization level (b) Average waiting time (c) Average en route time

Figure 6.3: Comparison of the two-direction call and no-call systems with patience level $T = 30$.

in Schaller (2017), which states that on-demand taxis has a higher proportion of unoccupied time between trips compared to yellow cabs in the "busy street" in New York City. In our analysis, we find that it occurs when the system utilization level is between 1 and 3 in the setting where passengers are less patient (and this region may shrink with the increase in the patience level). According to the data statistics by Castillo et al. (2018) (see Figure 1 in the paper), the Uber trip abandonment rates in Manhattan between December 2016 and May 2017 are mostly between 0.1 and 0.4, which suggests that the upper bound of the system utilization level ρ is between 1.11 and 1.67 (since the effective utilization level is always less than 1 in a stable system). Thus, our results provide support for the phenomenon observed in practice.

6.2 Idle Time Strategy

Idle time strategies in the existing literature of dispatching (e.g. McLeod 1972) concern how taxis proceed as soon as they drop off a passenger. In this section, we explore different idle time strategies in the call mechanism. In particular, we consider the following idle time strategies for the call mechanism: (i) The taxi keeps running when it finishes service (as in the base model); (ii) The taxi immediately stops when it finishes

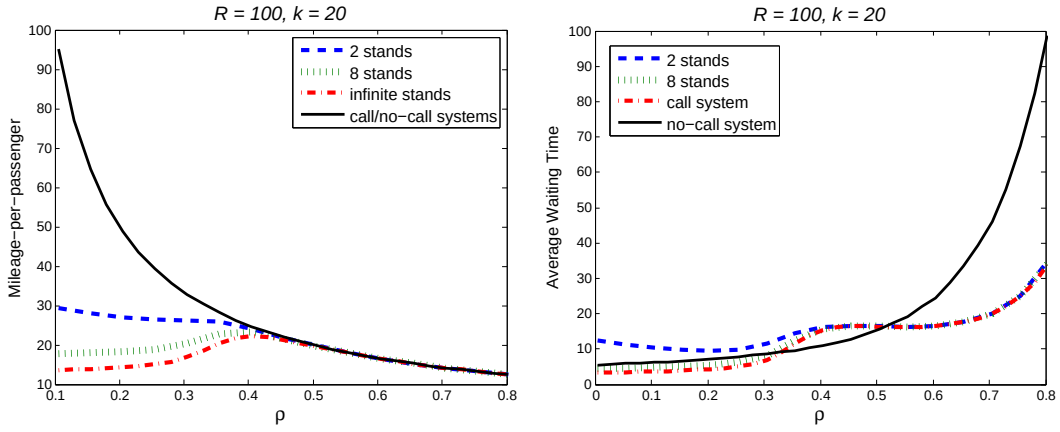
service until being requested again by the platform; and (iii) There are a few stands evenly distributed on the circular road, and the taxi drives toward the nearest stand when it finishes service and stays there until being requested again. In this setting, if a taxi is requested on the way to the nearest stand, it would respond to the request immediately. Since taxis in the no-call mechanism need to cruise in the system to pick up potential passengers, we assume that they would not stop at those taxi stands.

Let s denote the number of stands in the call mechanism. When $s = \infty$, every place can be treated as a taxi stand and thus taxis would just stay at the place where it becomes idle. When $s = 0$, a taxi would keep running even when it is unmatched. To measure the performance of the system, we define the *mileage-per-passenger* as the total distance traveled by all taxis over a long period of time divided by the total number of passengers served over the same period of time. A lower mileage-per-passenger would be favorable since it indicates a lower cost to serve a passenger. Figure 6.4 illustrates how the average mileage-per-passenger and the average waiting time change with ρ for call and no-call systems under different idle time strategies.

In particular, Figure 6.4(a) shows that when the system utilization level is low, compared to call/no-call systems with no stands, even a small number of taxi stands can significantly reduce the mileage-per-passenger by reducing the idle driving distance. In addition, the reduction of mileage per passenger increases in the number of stands s. However, when the system utilization level is high, these stands become useless because taxis would be fully occupied by ride requests and the idle time approaches zero.

Figure 6.4(b) compares the average waiting time in the call and no-call systems with different number of stands. It shows that the average waiting time in a call system decreases in the number of stands s, and the average waiting time in the call system with infinite stands ($s = \infty$) is very close to the average waiting time in the call system with no stands. Compared to a no-call system, when the system utilization level is

low, a call system with a small number of stands (e.g., $s = 2$) may have longer average waiting time. This is because when the system utilization level is low, it is very likely that a requested taxi is found idle at one of the few stands. In such a case, idle taxis aggregate at the limited locations instead of spreading out on the road, resulting in longer en route time compared to the call/no-call systems without stands. However, when the system utilization level is relatively high, stands becomes unoccupied and thus plays no role in reducing average waiting time.



(a) Mileage-per-passenger (b) Average waiting time

Figure 6.4: Comparison of the no-call and call systems with different number of taxi stands.

In sum, the operational value of taxi stands on either reducing average waiting time or running mileage diminishes when the system utilization level is relatively high. When the utilization is low, the installation of taxi stands could significantly reduce the idle driving mileage. Moreover, if the number of taxi stands is limited, then the mileage reduction is achieved at the expense of increasing passengers' average waiting time.

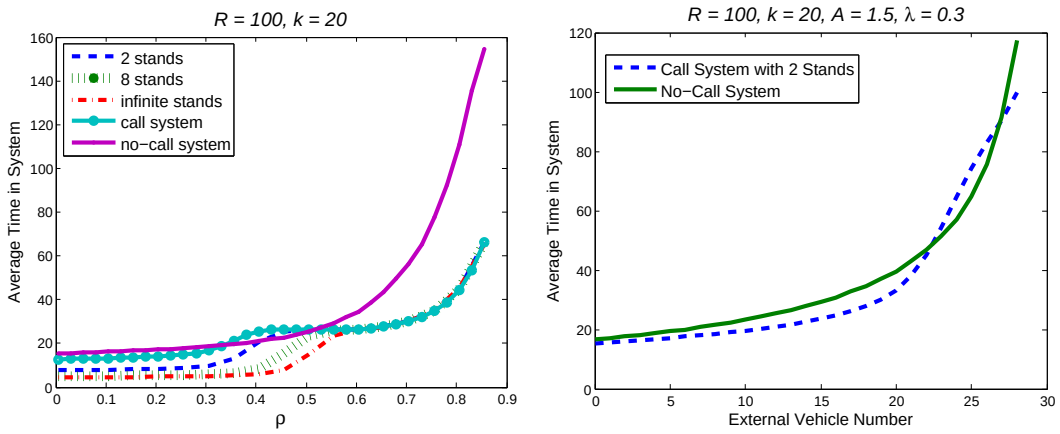
6.3 Traffic Congestion

When the number of vehicles in the system approaches the capacity limit of the road, congestion occurs and the speed of all vehicles decreases. In this section, we examine the impact of traffic congestion on the efficiency of call and no-call mechanisms. We assume that taxis in call and no-call systems follow the same idle time strategies as discussed in Section 6.2. That is, idle taxis in the no-call system keep running to search for passengers when completing rides, while idle taxis in the call system would be allowed to stay at stands (if there is any). As can be expected, the call mechanism may have the advantages of reducing the driving distance and thus mitigating traffic congestion.

To model traffic congestion, we adopt a linear model proposed by Jain and Smith (1997). Let l be the number of vehicles currently driving on the road, which includes both taxis and external vehicles in the system. In the linear model, the travel speed $V(l)$ linearly decreases in l, i.e., $V(l) = \frac{A}{C}(C + 1 - l)$, where A is the travel speed with only one vehicle on the road and C is the capacity of the road. To be consistent with the assumption made before, we let $V(k) = v$ by setting $C = \frac{A(k-1)}{A-v}$. Thus, the linear congestion model is reduced to $V(l) = \frac{kA-v}{k-1} - \frac{(A-v)}{k-1}l$, which satisfies $V(k) = v$ and $V(1) = A$.

Noting that the average service time is no longer a constant in such setting because the travel speed changes with the number of vehicles on the road due to traffic congestion. Therefore, to measure the performance of different systems, we use the notion of *average throughput time* in this section. The average throughput time is defined as the sum of the average waiting time and the average service time. In Figure 6.5(a), we illustrate the impact of stand numbers on the average throughput time. In the simulation, we take $A = 3$, and assume that the requested ride distance is exponentially distributed with mean $d = 10$. Recall that the call system with few stands ($s = 2$) may result in

longer average waiting time than the no-call system (see Figure 6.4b). However, when we take into account the impact of traffic congestion on travel speed, even a call system with two stands would result in shorter average throughput time than the no-call system when the system utilization level is not too large. When the system utilization level is large enough, all taxis would keep running on the road without stopping at stands. Thus the travel speed would be reduced to the minimum, and the average throughput time in the call system with stands is approaching to that without stands.



(a) (b)

Figure 6.5: The two-direction call/no-call systems with taxi stands and traffic congestion.

We next examine the impact of external traffic congestion on the performance of the call and no-call mechanisms. Figure 6.5(b) compares the average throughput time in the call and no-call systems when the number of external vehicles, denoted by n_e, changes. We set $A = 1.5$ and fix the arrival rate at $\lambda = 0.3$. To ensure that the travel speed $V(l)$ is always positive, the maximum vehicle capacity of the transportation network is $l_{max} = 58$. We compare the average throughput time in the call system with two stands with that in the no-call system. Figure 6.5(b) shows that the average throughput time in both systems increases with the number of external vehicles due to the deterioration

of the travel speed. When the number of external vehicles is relatively small, the call system with stands has a shorter average throughput time because of its advantage in reducing traffic congestion. However, with the increase of external vehicles, the advantage dissipates as the en route time becomes significantly large when the driving speed decreases, resulting in longer average throughput time. When the number of external vehicles continues to grow, both systems experience serious traffic congestion but meanwhile the en route time in the call system is reduced because of the increase of system utilization. Therefore, we observe that the call system has a shorter throughput time than the no-call system when the number of external vehicles is large.

Chapter 7

Conclusion and Discussion

The emergence of on-demand ride-hailing platforms over the past few years has greatly simplified the process of requesting transportation services. Although the on-demand platform has potential advantages in directing drivers to passengers in a timely manner, the platform also has its own challenges: A driver who commits to serve a passenger may miss the chance to serve another incoming passenger who is a shorter distance away. In this dissertation, we examine the performance of on-demand hailing and traditional street-hailing systems using a stylized model and $M/M/k$ queueing approximations. We find that while the average waiting time under the no-call mechanism always increases with system utilization as expected, the average waiting time under the call mechanism can decrease and then increase due to the non-monotonicity of the average en route time.

Distinct from standard queueing models, en route time plays an important role in the matching efficiency of a ride-hailing system. When passengers do not abandon the services, en route time in the call mechanism reaches the highest when the system is balanced, i.e., when the system utilization level is neither too high nor too low. In that case, taxis in a call mechanism spend much time in picking up passengers, during

which a driver may miss the chance to serve another incoming passenger who is at a shorter distance in the call mechanism. When the system utilization level is very high (with many waiting passengers) or very low (with many idle drivers), the average en route time is low. In the one direction systems, the average waiting time under the call mechanism is always higher than that under the no-call mechanism. This is because the no-call mechanism avoids forgoing possible future matching opportunities by postponing the matching between a driver and a passenger to as late as possible. However, the call mechanism can be more efficient than the no-call mechanism in the two-direction system despite of the disadvantages of forgoing possible future matching opportunities, because the call mechanism informs the taxi of the shortest distance to pick up a passenger. When the system utilization level is medium, the no-call mechanism may outperform the call mechanism in terms of average waiting time. The same results can be extended to a more complex grid network, where the call system results in higher waiting time than the no-call system when the utilization level is in the middle range, the grid network is not very complex, and the density of taxis in the system is high.

Based on the understanding of the two matching mechanisms, we sought to address the matching inefficiency that arises with the on-demand hailing platform by proposing a distance cap in responding to requests from passengers. The distance cap, on the one hand, reserves the advantages of the on-demand hailing system, while on the other hand helps in limiting the possibility of serving an incoming passenger. We propose a heuristic way to calculate the distance cap that not only is easily implementable but also effectively reduces average waiting time in the on-demand hailing system.

When passengers have limited patience in waiting and may abandon the services, we find that the call mechanism may result in higher average waiting time and lower effective utilization when the system utilization level is medium, which suggests that the call mechanism could be inefficient from both the perspectives of passengers and

the drivers.

We also examine several alternative idle time strategies in the call mechanism. We find that taxi stand is not necessarily useful when the system utilization level is relatively high. When the system utilization level is not too high, the installation of taxi stands could significantly reduce the mileage per passenger, but may result in higher average waiting time unless the number of stands is sufficient.

Last but not least, we find that traffic congestion could affect the performance of call and no-call systems. While the call system with a small number of stands has the advantage of reducing traffic congestion, the average throughput time in the call system could still be higher than the no-call system when the traffic congestion is in the medium range due to the longer en route time as a consequence of the traffic congestion.

In sum, our analysis shows that the on-demand hailing system is more efficient than the traditional street-hailing system in some circumstances while it is less efficient in others. This dissertation provides guidance for evaluating the efficiency of each system in different circumstances. Awareness of the advantages and disadvantages of the on-demand hailing and street-hailing systems would not only assist policy makers evaluate the potential outcome of adopting on-demand hailing or street-hailing but also help the platform improve the matching efficiency in some circumstances. Our study may provide justifications for the New York City government to allow a certain number of taxis dedicated to street-hailing through Street Hail Livery (SHL) since we have shown that street-hailing has its own value in operations and should be reserved under certain conditions. The results in the dissertation also verify the empirical finding in Schaller (2017) that the unoccupied time of the on-demand ride-hailing taxis in New York could be higher than the street-hailing taxis. There are many future research directions following this work, such as increasing the accuracy of the approximation scheme, extending the theoretical results to more complex cases, and considering endogenous price, wage

and capacity decisions in the ride-hailing systems that would affect the demand and labor supply and ultimately the efficiency of the system.

References

P. Afèche, Z. Liu, and C. Maglaras. Ride-hailing networks with strategic drivers: The impact of platform control capabilities on performance. 2018. Available at SSRN: https://ssrn.com/abstract=3120544.

M. Akbarpour, S. Li, and S. Shayan. Thickness and information in dynamic matching markets. Available at SSRN:https://ssrn.com/abstract=2394319, 2016.

G. Allon, A. Bassamboo, and B. Çil. Large-scale service marketplaces: The role of the moderating firm. *Management Science*, 58(10):1854–1872, 2012.

R. Anderson, I. Ashlagi, D. Gamarnik, and Y. Kanoria. A dynamic model of barter exchange. In *In SODA'15: Proceedings of the 26th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1925–1933, 2015.

M. Baccara, S. Lee, and L. Yariv. Optimal dynamic matching. Available at SSRN:https://ssrn.com/abstract=2641670, 2015.

W. Bailey and T. Clark. Taxi management and route control: A systems study and simulation experiment. In *Proceedings of the 24th Conference on Winter Simulation*, pages 1217–1222, 1992.

M. Balinski and R. Gomory. A primal method for the assignment and transportation problems. *Management Science*, 10(3):578–593, 1964.

R. Barr, F. Glover, and D. Klingman. The alternating basis algorithm for assignment problems. *Mathematical Programming*, 13(1):1–13, 1977.

S. Benjaafar, H. Bernhard, and C. Courcoubetis. Drivers, riders and service providers: The impact of the sharing economy on mobility. 2017.

S. Benjaafar, J.-Y. Ding, G. Kong, and T. Taylor. Labor welfare in on-demand service platforms. 2018a.

S. Benjaafar, G. Kong, X. Li, and C. Courcoubetis. Peer-to-peer product sharing: Implications for ownership, usage and social welfare in the sharing economy. *Management Science*, 2018b. ePub ahead of print May 2018, https://pubsonline.informs.org/doi/pdf/10.1287/mnsc.2017.2970.

D. Bertsekas. A new algorithm for the assignment problem. *Mathematical Programming*, 21(1):152–171, 1981.

K. Bimpikis, O. Candogan, and S. Daniela. Spatial pricing in ride-sharing networks. 2016.

O. Boxma, Q. Deng, and A. Zwart. Waiting-time asymptotics for the m/g/2 queue with heterogeneous servers. *Queueing Systems*, 40(1):5–31, 2002.

P. Cachon, M. Daniels, and R. Lobel. The role of surge pricing on a service platform with self-scheduling capacity. *Manufacturing & Service Operations Management*, 19 (3):368384, 2017.

J. Castillo, D. Knoepfle, and G. Weyl. Surge pricing solves the wild goose chase. In *Proceedings of the 2017 ACM Conference on Economics and Computation*, pages 241–242. ACM, 2017.

F. Castro, O. Besbes, and I. Lobel. Surge pricing and its spatial supply response. 2018.

E. Coffman and E. Gilbert. A continuous polling system with constant service times. *IEEE Transactions on Information Theory*, 32(4):584–591, 1986.

Z. Cullen and C. Farronato. Outsourcing tasks online: Matching supply and demand on peer-to-peer internet platforms. Working Paper, 2014.

G. Dantzig. *Linear programming and extensions*. Princeton university press, 2016.

Y. Fawaz and F. Newell. Optimal spacings for a rectangular grid transportation network: A hierarchy structure. *Transportation Research*, 10(2):111–119, 1976.

P. Fraiberger and A. Sundararajan. Peer-to-peer rental markets in the sharing economy. Available at SSRN: https://ssrn.com/abstract=2574337, 2017.

M. Gerrard. Comparison of taxi and dial-a-bus services. *Transportation Science*, 8(2): 85–101, 1974.

H. Guda and U. Subramanian. Your uber is arriving: Managing on-demand workers through surge pricing, forecast communication and worker incentives. *Management Science*, 2018. Forthcoming.

V. Gupta, M. Harchol-Balter, J. Dai, and B. Zwart. On the inapproximability of m/g/k: why two moments of job size distribution are not enough. *Queueing Systems*, 64(1): 5–48, 2010.

I. Gurvich, M. Lariviere, and A. Moreno. Operations in the on-demand economy: Staffing services with self-scheduling capacity. Working Paper, 2015.

J. Hall, J. Horton, and D. Knoepfle. Labor market equilibration: Evidence from uber. Technical report, Working Paper, 1–42, 2017.

L. Hall, A. Schulz, D. Shmoys, and J. Wein. Scheduling to minimize average completion time: Off-line and on-line approximation algorithms. *Mathematics of operations research*, 22(3):513–544, 1997.

J. Hawkins. It took Uber five years to get to a billion rides, and its Chinese rival just did it in one. *The Verge*, 2016a.

J. Hawkins. Uber just completed its two-billionth trip @Verge. `http://www.theverge.com/2016/7/18/12211710/uber-two-billion-trip-announced-kalanick-china-didi`, 2016b. Accessed: 2017-04-21.

J. Hoogeveen and P. Vestjens. A best possible deterministic on-line algorithm for minimizing maximum delivery time on a single machine. *SIAM Journal on Discrete Mathematics*, 13(1):56–63, 2000.

M. Hu and Y. Zhou. Dynamic type matching. Available at SSRN: https://ssrn.com/abstract=2592622, 2016.

M. Hung. A polynomial simplex method for the assignment problem. *Operations Research*, 31(3):595–600, 1983.

R. Jain and M. Smith. Modeling vehicular traffic flow using M/G/C/C state dependent queueing models. *Transportation Science*, 31(4):324–336, 1997.

B. Jiang and L. Tian. Collaborative consumption: Strategic and economic implications of product sharing. *Management Science*, 2016. ePub ahead of print November 16, http://dx.doi.org/10.1287/mnsc.2016.2647.

R. Jonker and A. Volgenant. A shortest augmenting path algorithm for dense and sparse linear assignment problems. *Computing*, 38(4):325–340, 1987.

D. P. Kroese and V. Schmidt. A continuous polling system with general service times. *The Annals of Applied Probability*, pages 906–927, 1992.

M. McLeod. The operation and performance of a taxi fleet. Master's thesis, Massachusetts Institute of Technology, 1972.

R. Meyer and H. Wolfe. The organization and operation of a taxi fleet. *Naval Research Logistics Quarterly*, 8(2):137–150, 1961.

K. G. Murty. *Network programming*. Prentice-Hall, Inc., 1992.

E. Ozkan and A. Ward. Dynamic matching for real-time ridesharing. Available at SSRN: https://ssrn.com/abstract=2844451, 2017.

M. Pinedo. *Scheduling: theory, algorithms, and systems*. Springer, 2016.

C. Riquelme, S. Banerjee, and R. Johari. Pricing in ride-share platforms: A queueing-theoretic approach. Available at SSRN:https://ssrn.com/abstract=2568258, 2015.

B. Rogers. The social costs of Uber. *University of Chicago Law Review Dialogue*, 82: 85–104, 2015.

S. Sani and O. A. Daman. The m/g/2 queue with heterogeneous servers under a controlled service discipline: Stationary performance analysis. *International Journal of Applied Mathematics*, 45(1), 2015.

B. Schaller. Empty seats, full streets: Fixing manhattan's traffic problem. Available at http://www.schallerconsult.com/rideservices/emptyseats.pdf, 2017.

D. Shmoys, J. Wein, and D. Williamson. Scheduling parallel machines on-line. *SIAM journal on computing*, 24(6):1313–1331, 1995.

Z. Spivey and B. Powell. The dynamic assignment problem. *Transportation Science*, 38 (4):399–419, 2004.

J. Steinberg. Smartphone taxi e-hail apps: New convenience or potential deathtrap. *The Forbes*, 2012.

P. Sweeney. Shanghai government cracks down on taxi booking apps. *Business Insider*, 2014. (Feb 26), http://www.businessinsider.com/r-shanghai-government-cracks-down-on-taxi-booking-apps-2014-26.

S. Tang, J. Bai, C. So, X. Chen, and H. Wang. Coordinating supply and demand on an on-demand platform: Price, wage, and payout ratio. Working Paper, 2016.

T. Taylor. On-demand service platforms. Working Paper, 2016.

N. Tomizawa. On some techniques useful for solution of transportation network problems. *Networks*, 1(2):173–194, 1971.

P. Welch. The statistical analysis of simulation results. In *The Computer Performance Modeling Handbook*, pages 268–328. Academic Press, 1983.

D. Wood. *The Computation of Polylogarithms*. University of Kent at Canterbury, 1992.

Appendix A

Proofs

A.1 Proof of Lemma A.1.1

Lemma A.1.1 *Consider an $M/M/k$ queue with arrival rate λ and state-dependent service rate*

$$\mu_n = \begin{cases} \frac{n(k-n+2)}{R}, & \text{if } n \leq k, \\ \frac{k(n-k+1)}{R}, & \text{if } n > k. \end{cases}$$

Let $\overline{\mathcal{W}}(\lambda, k, R)$ be the average waiting time. Then for any $R > 0$ and $k \geq 2$, there exists an arrival rate $\lambda^*(k, R)$ such that $\frac{\partial \overline{\mathcal{W}}(\lambda,k,R)}{\partial \lambda}\Big|_{\lambda=\lambda^*(k,R)} < 0$.

Proof of Lemma A.1.1. By standard queueing theory, $\overline{\mathcal{W}}(\lambda, k, R) = \frac{\sum_{i=1}^{\infty} \frac{i\lambda^{i-1}}{\mu_1\mu_2\cdots\mu_i}}{1+\sum_{i=1}^{\infty} \frac{\lambda^i}{\mu_1\mu_2\cdots\mu_i}}$.

For $\sum_{i=1}^{\infty}\frac{\lambda^i}{\mu_1\mu_2\cdots\mu_i}$, we have $\sum_{i=1}^{\infty}\frac{\lambda^i}{\mu_1\mu_2\cdots\mu_i}=\sum_{i=1}^{k-1}\frac{\lambda^i}{\mu_1\mu_2\cdots\mu_i}+\sum_{i=k}^{\infty}\frac{\lambda^i}{\mu_1\mu_2\cdots\mu_i}$, where

$$\begin{aligned}
\sum_{i=k}^{\infty}\frac{\lambda^i}{\mu_1\mu_2\cdots\mu_i} &= \frac{\lambda^k}{\mu_1\mu_2\cdots\mu_k}\left(1+\sum_{i=k+1}^{\infty}\frac{\lambda^{i-k}}{\mu_{k+1}\cdots\mu_i}\right)\\
&= \frac{\lambda^k}{\mu_1\mu_2\cdots\mu_k}\left(1+\sum_{i=k+1}^{\infty}\frac{(\lambda R/k)^{i-k}}{(i-k+1)!}\right)\\
&= \frac{\lambda^k}{\mu_1\mu_2\cdots\mu_k}\frac{k}{\lambda R}\left((\lambda R/k)+\sum_{i=k+1}^{\infty}\frac{(\lambda R/k)^{i-k+1}}{(i-k+1)!}\right)\\
&= \frac{\lambda^{k-1}}{2\mu_1\mu_2\cdots\mu_{k-1}}\sum_{i=1}^{\infty}\frac{(\lambda R/k)^i}{i!}=\frac{\lambda^{k-1}}{2\mu_1\mu_2\cdots\mu_{k-1}}(e^{\lambda R/k}-1).
\end{aligned}$$

Similarly, for $\sum_{i=1}^{\infty} i\frac{\lambda^{i-1}}{\mu_1\mu_2\cdots\mu_i}$, we have $\sum_{i=1}^{\infty} i\frac{\lambda^{i-1}}{\mu_1\mu_2\cdots\mu_i}=\sum_{i=1}^{k-1} i\frac{\lambda^{i-1}}{\mu_1\mu_2\cdots\mu_i}+\sum_{i=k}^{\infty} i\frac{\lambda^{i-1}}{\mu_1\mu_2\cdots\mu_i}$, where

$$\begin{aligned}
\sum_{i=k}^{\infty} i\frac{\lambda^{i-1}}{\mu_1\mu_2\cdots\mu_i} &= \frac{\lambda^{k-1}}{\mu_1\mu_2\cdots\mu_k}\left(k+\sum_{i=k+1}^{\infty} i\frac{\lambda^{i-k}}{\mu_{k+1}\cdots\mu_i}\right)\\
&= \frac{\lambda^{k-1}}{\mu_1\mu_2\cdots\mu_k}\frac{k}{\lambda R}\left(k\frac{\lambda R}{k}+\sum_{i=k+1}^{\infty} i\frac{(\lambda R/k)^{i-k+1}}{(i-k+1)!}\right)\\
&= \frac{\lambda^{k-2}}{2\mu_1\mu_2\cdots\mu_{k-1}}\sum_{i=k}^{\infty} i\frac{(\lambda R/k)^{i-k+1}}{(i-k+1)!}\\
&= \frac{\lambda^{k-2}}{2\mu_1\mu_2\cdots\mu_{k-1}}\left((k-1)\sum_{i=k}^{\infty}\frac{(\lambda R/k)^{i-k+1}}{(i-k+1)!}+\sum_{j=1}^{\infty}\frac{j(\lambda R/k)^j}{j!}\right)\\
&= \frac{\lambda^{k-2}}{2\mu_1\mu_2\cdots\mu_{k-1}}\left[(k-1)(e^{\lambda R/k}-1)+\frac{\lambda R}{k}e^{\lambda R/k}\right].
\end{aligned}$$

Therefore

$$\overline{\mathcal{W}}(\lambda,k,R)=\frac{\sum_{i=1}^{k-1} i\frac{\lambda^{i-1}}{\mu_1\mu_2\cdots\mu_i}+\frac{\lambda^{k-2}}{2\mu_1\mu_2\cdots\mu_{k-1}}[(k-1)(e^{\lambda R/k}-1)+\frac{\lambda R}{k}e^{\lambda R/k}]}{1+\sum_{i=1}^{k-1}\frac{\lambda^i}{\mu_1\mu_2\cdots\mu_i}+\frac{\lambda^{k-1}}{2\mu_1\mu_2\cdots\mu_{k-1}}(e^{\lambda R/k}-1)}.$$

For any given R and $k\geq 2$, we note that $\overline{\mathcal{W}}\big|_{\lambda=0}=R/(k+1)$ and $\overline{\mathcal{W}}\big|_{\lambda\to\infty}=R/k$.

Also we claim that for any given R and $k \geq 2$, there exists $\tilde{\lambda}$ such that $\overline{\mathcal{W}}\big|_{\lambda=\tilde{\lambda}} > R/k$. To see this, when $k = 2$, we can simplify the expression of $\overline{\mathcal{W}}$ and get that $\overline{\mathcal{W}} = \frac{1+(1+\lambda R/2)\exp(\lambda R/2)}{\lambda+6/R+\lambda\exp(\lambda R/2)}$. By taking $\tilde{\lambda}$ satisfying $\exp(\tilde{\lambda}R/2) > 2 + \tilde{\lambda}R/2$, we have $\overline{\mathcal{W}}\big|_{\lambda=\tilde{\lambda}} > R/2$. When $k > 2$, we have

$$\begin{aligned}\overline{\mathcal{W}} &> \frac{\frac{\lambda^{k-2}}{2\mu_1\mu_2\cdots\mu_{k-1}}[(k-1)(e^{\lambda R/k}-1)+\frac{\lambda R}{k}e^{\lambda R/k}]}{1+\sum_{i=1}^{k-1}\frac{\lambda^i}{\mu_1\mu_2\cdots\mu_i}+\frac{\lambda^{k-1}}{2\mu_1\mu_2\cdots\mu_{k-1}}(e^{\lambda R/k}-1)} \\ &> \frac{\frac{\lambda^{k-1}}{2\mu_1\mu_2\cdots\mu_{k-1}}(k-1+e^{\lambda R/k})\frac{R}{k}}{1+\sum_{i=1}^{k-1}\frac{\lambda^i}{\mu_1\mu_2\cdots\mu_i}+\frac{\lambda^{k-1}}{2\mu_1\mu_2\cdots\mu_{k-1}}(e^{\lambda R/k}-1)}.\end{aligned}$$

Taking $\tilde{\lambda}$ satisfying $\frac{k\tilde{\lambda}^{k-1}}{2\mu_1\mu_2\cdots\mu_{k-1}} > 1+\sum_{i=1}^{k-1}\frac{\tilde{\lambda}^i}{\mu_1\mu_2\cdots\mu_i}$ (note such $\tilde{\lambda}$ must exist since the left-hand side is greater than the right-hand side when $\tilde{\lambda}$ is large), we obtain $\overline{\mathcal{W}}\big|_{\lambda=\tilde{\lambda}} > R/k$.

Thus for any given R and $k \geq 2$, $\overline{\mathcal{W}}\big|_{\lambda=0} = R/(k+1)$ and $\overline{\mathcal{W}}\big|_{\lambda\to\infty} = R/k$, and there exists $\tilde{\lambda}$ such that $\overline{\mathcal{W}}\big|_{\lambda=\tilde{\lambda}} > R/k$. Also it is easy to see that $\overline{\mathcal{W}}$ is continuously differentiable in λ on $\lambda > 0$. Therefore there exists $\lambda^*(k,R)$ such that $\frac{\partial\overline{\mathcal{W}}(\lambda,k,R)}{\partial\lambda}\Big|_{\lambda=\lambda^*(k,R)} < 0$.
□

A.2 Proof of Theorem 3.3.1

Proof of Theorem 3.3.1. In the following, we will prove the five parts of Theorem 3.3.1 separately. For the ease of notation, we omit the superscripts in μ_is, λ_is and $\mathcal{W}(\cdot)$ as the meanings are clear from the context.

Part 1. For the approximated one-direction call system, the expected waiting time $\mathcal{W}(\lambda, d, v, k, R)$ satisfies that $\mathcal{W} = \frac{\sum_{i=1}^{\infty}\frac{i\lambda^{i-1}}{\mu_1\mu_2\cdots\mu_i}}{1+\sum_{i=1}^{\infty}\frac{\lambda^i}{\mu_1\mu_2\cdots\mu_i}} - d/v$, where μ_is are defined in (A.4). Define $\mu = v/d$. When $\lambda \geq k\mu$, since $\mu_i \leq k\mu$ for all i, it is easy to see that $\mathcal{W}(\lambda, d, v, k, R) = \infty$. Therefore it remains to prove that when $\lambda < k\mu$, $\mathcal{W}(\lambda, d, v, k, R) < \infty$. It is equivalent to show $\sum_{i=1}^{\infty}\frac{i\lambda^{i-1}}{\mu_1\mu_2\cdots\mu_i}$ and $\sum_{i=1}^{\infty}\frac{\lambda^i}{\mu_1\mu_2\cdots\mu_i}$ are

both finite when $0 \leq \lambda < k\mu$. Note that

$$\begin{aligned}\sum_{i=1}^{\infty} \frac{\lambda^i}{\mu_1\mu_2\cdots\mu_i} &= \sum_{i=1}^{N} \frac{\lambda^i}{\mu_1\mu_2\cdots\mu_i} + \frac{\lambda^N}{\mu_1\mu_2\cdots\mu_N} \sum_{i=1}^{\infty} \frac{\lambda^i}{\mu_{N+1}\mu_{N+2}\cdots\mu_{N+i}} \\ &< \sum_{i=1}^{N} \frac{\lambda^i}{\mu_1\mu_2\cdots\mu_i} + \frac{\lambda^N}{\mu_1\mu_2\cdots\mu_N} \sum_{i=1}^{\infty} \frac{\lambda^i(1/\mu+\epsilon)^i}{k^i} \\ &< \sum_{i=1}^{N} \frac{\lambda^i}{\mu_1\mu_2\cdots\mu_i} + \frac{\lambda^N}{\mu_1\mu_2\cdots\mu_N} \frac{k\mu+\lambda}{k\mu-\lambda} < \infty,\end{aligned}$$

where the first inequality is because $\mu_l > \frac{k}{1/\mu+\epsilon}$ for any $l \geq N+1$. Similarly, we can also prove $\sum_{i=1}^{\infty} \frac{i\lambda^{i-1}}{\mu_1\mu_2\cdots\mu_i} < \infty$ when $0 \leq \lambda < k\mu$. Therefore, $\mathcal{W}(\lambda, d, v, k, R) < \infty$ if and only if $0 \leq \lambda < k\mu = kv/d$ and thus the first part is proved.

Part 2. Define $\mu = v/d$. We first show that $\frac{\partial \mathcal{W}}{\partial \lambda}\big|_{\lambda=0} \geq 0$. To see that, we take derivative of $\mathcal{W}$ with respect to λ, we have

$$\frac{\partial \mathcal{W}}{\partial \lambda} = \frac{\sum_{i=2}^{\infty} \frac{i(i-1)\lambda^{i-2}}{\mu_1\mu_2\cdots\mu_i}(1+\sum_{i=1}^{\infty} \frac{\lambda^i}{\mu_1\mu_2\cdots\mu_i}) - (\sum_{i=1}^{\infty} \frac{i\lambda^{i-1}}{\mu_1\mu_2\cdots\mu_i})^2}{(1+\sum_{i=1}^{\infty} \frac{\lambda^i}{\mu_1\mu_2\cdots\mu_i})^2}.$$

We have that

$$\left.\frac{\partial \mathcal{W}}{\partial \lambda}\right|_{\lambda=0} = \frac{2}{\mu_1\mu_2} - \frac{1}{\mu_1^2} = \begin{cases} \frac{1}{\mu_1}(\frac{1}{\mu} + \frac{R}{2v}), & \text{if } k = 1, \\ \frac{R}{6\mu_1 v}, & \text{if } k = 2, \\ \frac{1}{\mu_1}(\frac{R}{kv} - \frac{R}{(k+1)v}), & \text{if } k > 2. \end{cases}$$

Therefore, $\frac{\partial \mathcal{W}}{\partial \lambda}\big|_{\lambda=0} \geq 0$.

Next, we prove that $\underline{\lim}_{\lambda \to k\mu-} \frac{\partial \mathcal{W}(\lambda,d,v,k,R)}{\partial \lambda} > 0$. It suffices to show that

$$\underline{\lim}_{\lambda \to k\mu-} \left\{ \sum_{i=2}^{\infty} \frac{i(i-1)\lambda^{i-2}}{\mu_1\mu_2\cdots\mu_i} \left(1 + \sum_{i=1}^{\infty} \frac{\lambda^i}{\mu_1\mu_2\cdots\mu_i}\right) - \left(\sum_{i=1}^{\infty} \frac{i\lambda^{i-1}}{\mu_1\mu_2\cdots\mu_i}\right)^2 \right\} > 0. \quad \text{(A.1)}$$

Let $0 < \epsilon < 1$ be a constant such that $(1-\epsilon)^2(\mu R/v+2) - (1+\epsilon)^2(\mu R/v+1) > \frac{1}{2}$, and $M = \left\lceil \max\left\{ \frac{\mu R/v-k+1}{\sqrt[\mu R/v]{\frac{1}{1-\epsilon}}-1}, \frac{k-2}{1-\sqrt[\mu R/v]{\frac{1}{1+\epsilon}}} \right\} \right\rceil + 1$. Here we use $\lceil x \rceil$ to denote the smallest integer that is greater than x. In the following, for the ease of notation, we assume $\mu R/v$ is an integer. The case when $\mu R/v$ is not an integer can be proved in a very similar way, however, the notation would be much messier. For $i \geq \mu R/v + M + 1$, by the definition of μ_n, we have

$$\begin{aligned} \frac{\lambda^i}{\mu_1 \cdots \mu_i} &= \left(\frac{\lambda}{k\mu}\right)^i \frac{k^k}{k!} \prod_{\ell=1}^{k} \left(1 + \frac{\mu R}{(k-\ell+2)v}\right) \prod_{\ell=k+1}^{i} \left(1 + \frac{\mu R}{(\ell-k+1)v}\right) \\ &= C_1 \left(\frac{\lambda}{k\mu}\right)^i \prod_{\ell=M+1}^{i} \left(1 + \frac{\mu R}{(\ell-k+1)v}\right) \end{aligned}$$

where $C_1 = \frac{k^k}{k!} \prod_{\ell=1}^{k}(1 + \frac{\mu R}{(k-\ell+2)v}) \prod_{\ell=k+1}^{M}(1 + \frac{\mu R}{(\ell-k+1)v})$ is a constant. We have

$$\begin{aligned} & \prod_{\ell=M+1}^{i} \left(1 + \frac{\mu R}{(\ell-k+1)v}\right) \\ = \; & \frac{(\mu R/v + M + 2 - k)(\mu R/v + M + 3 - k) \cdots (\mu R/v + i + 1 - k)}{(M+2-k)(M+3-k)\cdots(i+1-k)} \\ = \; & \frac{(i+2-k)\cdots(i+1-k+\mu R/v)}{(M+2-k)\cdots(M+1-k+\mu R/v)} \\ \in \; & \left(\left(\frac{i+1-k+\mu R/v}{M+1-k+\mu R/v}\right)^{\mu R/v}, \left(\frac{i+2-k}{M+2-k}\right)^{\mu R/v} \right) \end{aligned}$$

where in the second equality $i \geq \mu R/v + M + 1$ and $\mu R/v \in \mathbb{Z}_+$ are applied. By the definition of M, we have that

$$\prod_{\ell=M+1}^{i} \left(1 + \frac{\mu R}{(\ell-k+1)v}\right) \in \left\{ (1-\epsilon)\left(\frac{i}{M}\right)^{\mu R/v}, (1+\epsilon)\left(\frac{i}{M}\right)^{\mu R/v} \right\}.$$

Therefore, for $i \geq \mu R/v + M + 1$,

$$\frac{\lambda^i}{\mu_1 \cdots \mu_i} \in \left\{ C_1(1-\epsilon)\left(\frac{\lambda}{k\mu}\right)^i \left(\frac{i}{M}\right)^{\mu R/v}, C_1(1+\epsilon)\left(\frac{\lambda}{k\mu}\right)^i \left(\frac{i}{M}\right)^{\mu R/v} \right\}.$$

Similarly, we have

$$\frac{i\lambda^{i-1}}{\mu_1 \cdots \mu_i} \in \left\{ \frac{C_1(1-\epsilon)i}{k\mu}\left(\frac{\lambda}{k\mu}\right)^{i-1} \left(\frac{i}{M}\right)^{\mu R/v}, \frac{C_1(1+\epsilon)i}{k\mu}\left(\frac{\lambda}{k\mu}\right)^{i-1} \left(\frac{i}{M}\right)^{\mu R/v} \right\}$$

and

$$\begin{aligned} &\frac{i(i-1)\lambda^{i-2}}{\mu_1 \cdots \mu_i} \\ \in\ &\left\{ \frac{C_1(1-\epsilon)i(i-1)}{(k\mu)^2}\left(\frac{\lambda}{k\mu}\right)^{i-2} \left(\frac{i}{M}\right)^{\mu R/v}, \frac{C_1(1+\epsilon)i(i-1)}{(k\mu)^2}\left(\frac{\lambda}{k\mu}\right)^{i-2} \left(\frac{i}{M}\right)^{\mu R/v} \right\}. \end{aligned}$$

Therefore, we have

$$\sum_{i=1}^{\infty} \frac{\lambda^i}{\mu_1 \cdots \mu_i} = \sum_{i=1}^{M} \frac{\lambda^i}{\mu_1 \cdots \mu_i} + \sum_{i=M+1}^{\infty} \frac{\lambda^i}{\mu_1 \cdots \mu_i} \geq C_2 + \frac{C_1(1-\epsilon)}{M^{\mu R/v}} \sum_{i=1}^{\infty} \left(\frac{\lambda}{k\mu}\right)^i i^{\mu R/v},$$

where $C_2 = -MC_1(1-\epsilon) \leq \sum_{i=1}^{M} \frac{\lambda^i}{\mu_1 \cdots \mu_i} - \frac{C_1(1-\epsilon)}{M^{\mu R/v}} \sum_{i=1}^{M} \left(\frac{\lambda}{k\mu}\right)^i i^{\mu R/v}$ is a constant. Similarly,

$$\sum_{i=2}^{\infty} \frac{i(i-1)\lambda^{i-2}}{\mu_1 \cdots \mu_i} \geq C_2' + \frac{C_1(1-\epsilon)}{(k\mu)^2 M^{\mu R/v}} \sum_{i=2}^{\infty} \left(\frac{\lambda}{k\mu}\right)^{i-2} (i-1) i^{\mu R/v+1},$$

where

$$C_2' = -C_1 M^3(1-\epsilon)/(k\mu)^2 \leq \sum_{i=2}^{M} \frac{i(i-1)\lambda^{i-2}}{\mu_1 \cdots \mu_i} - \frac{C_1(1-\epsilon)}{(k\mu)^2 M^{\mu R/v}} \sum_{i=2}^{M} (\frac{\lambda}{k\mu})^{i-2} (i-1) i^{\mu R/v+1}$$

and

$$\sum_{i=1}^{\infty} \frac{i\lambda^{i-1}}{\mu_1 \cdots \mu_i} \leq C_2'' + \frac{C_1(1-\epsilon)}{k\mu M^{\mu R/v}} \sum_{i=1}^{\infty} \left(\frac{\lambda}{k\mu}\right)^{i-1} i^{\mu R/v+1},$$

with $C_2'' = M^2 k^M \geq \sum_{i=1}^{M} \frac{i\lambda^{i-1}}{\mu_1 \cdots \mu_i} - \frac{C_1(1-\epsilon)}{k\mu M^{\mu R}} \sum_{i=1}^{M} (\frac{\lambda}{k\mu})^{i-1} i^{\mu R/v+1}$.

In the following, we define

$$f_0(p) = \sum_{i=1}^{\infty} p^i i^{\mu R/v}, \quad f_1(p) = \sum_{i=1}^{\infty} p^{i-1} i^{\mu R/v+1} \quad \text{and} \quad f_2(p) = \sum_{i=2}^{\infty} p^{i-2}(i-1) i^{\mu R/v+1}.$$

Using these functions, the left-hand side of (A.1) can be written as:

$$\begin{aligned} & \textstyle\sum_{i=2}^{\infty} \frac{i(i-1)\lambda^{i-2}}{\mu_1\mu_2\cdots\mu_i} \left(1 + \sum_{i=1}^{\infty} \frac{\lambda^i}{\mu_1\mu_2\cdots\mu_i}\right) - \left(\sum_{i=1}^{\infty} \frac{i\lambda^{i-1}}{\mu_1\mu_2\cdots\mu_i}\right)^2 \\ \geq \; & \left(C_2' + \frac{C_1(1-\epsilon)}{(k\mu)^2 M^{\mu R/v}} f_2\left(\frac{\lambda}{k\mu}\right)\right)\left(C_2 + \frac{C_1(1-\epsilon)}{M^{\mu R/v}} f_0\left(\frac{\lambda}{k\mu}\right)\right) - \left(C_2'' + \frac{C_1(1+\epsilon)}{k\mu M^{\mu R/v}} f_1\left(\frac{\lambda}{k\mu}\right)\right)^2. \end{aligned} \tag{A.2}$$

In the following, we use the notation of polylogarithm, in which $\mathrm{Li}_s(z) = \sum_{i=1}^{\infty} \frac{z^i}{i^s}$. Then we have

$$f_0(p) = \mathrm{Li}_{-\mu R/v}(p), \quad f_1(p) = \frac{1}{p}\mathrm{Li}_{-\mu R/v-1}(p),$$

and $f_2(p) = \frac{1}{p^2}(\mathrm{Li}_{-\mu R/v-2}(p) - \mathrm{Li}_{-\mu R/v-1}(p))$. By the limiting behavior of the polylogarithm (Wood 1992), we have that

$$\lim_{p\to 1^-} f_0(p) = \frac{\Gamma(1+\mu R/v)}{(-\log p)^{\mu R/v+1}}, \quad \lim_{p\to 1^-} f_1(p) = \frac{\Gamma(2+\mu R/v)}{(-\log p)^{\mu R/v+2}},$$

and $\lim_{p\to 1^-} f_2(p) = \frac{\Gamma(3+\mu R/v)}{(-\log p)^{\mu R/v+3}} - \frac{\Gamma(2+\mu R/v)}{(-\log p)^{\mu R/v+2}}$, where $\Gamma(\cdot)$ is the Gamma function. Therefore,

$$\begin{aligned}
& \lim_{p\to 1^-}(1-\epsilon)^2 f_0(p) f_2(p) - (1+\epsilon)^2 f_1(p)^2 \\
= \; & ((1-\epsilon)^2(\mu R/v)!(\mu R/v+2)! - (1+\epsilon)^2(\mu R/v+1)!^2)(-\log p)^{-2\mu R/v-4} \\
& -(1-\epsilon)^2(\mu R/v)!(\mu R/v+1)!(-\log p)^{-2\mu R/v-3} \\
= \; & (\mu R/v)!(\mu R/v+1)![(1-\epsilon)^2(\mu R/v+2) - (1+\epsilon)^2(\mu R/v+1)](-\log p)^{-2\mu R/v-4} \\
& -(1-\epsilon)^2(\mu R/v)!(\mu R/v+1)!(-\log p)^{-2\mu R/v-3} \\
\geq \; & (\mu R/v)!(\mu R/v+1)!\left(\frac{1}{2}(-\log p)^{-2\mu R/v-4} - (1-\epsilon)^2(-\log p)^{-2\mu R/v-3}\right) \\
= \; & \mathcal{O}(-\log p)^{-2\mu R/v-4} > 0
\end{aligned}$$

where the inequality is because of the definition of ϵ. Based on these results, for equation (A.2), we further have that

$$\begin{aligned}
& \underline{\lim}_{\lambda\to k\mu-}\left(\sum_{i=2}^{\infty}\frac{i(i-1)\lambda^{i-2}}{\mu_1\mu_2\cdots\mu_i}\right)\left(1+\sum_{i=1}^{\infty}\frac{\lambda^i}{\mu_1\mu_2\cdots\mu_i}\right) - \left(\sum_{i=1}^{\infty}\frac{i\lambda^{i-1}}{\mu_1\mu_2\cdots\mu_i}\right)^2 \\
\geq \; & \underline{\lim}_{\lambda\to k\mu-}\left(C_2'' + \frac{C_1(1-\epsilon)}{(k\mu)^2 M^{\mu R/v}} f_2\left(\frac{\lambda}{k\mu}\right)\right)\left(C_2 + \frac{C_1(1-\epsilon)}{M^{\mu R/v}} f_0\left(\frac{\lambda}{k\mu}\right)\right) \\
- \; & \left(C_2' + \frac{C_1(1+\epsilon)}{k\mu M^{\mu R/v}} f_1\left(\frac{\lambda}{k\mu}\right)\right)^2 \\
= \; & \mathcal{O}(-\log(\tfrac{\lambda}{k\mu}))^{-2\mu R/v-4}.
\end{aligned} \tag{A.3}$$

The last step is because the other terms are smaller in order than $\mathcal{O}(-\log(\frac{\lambda}{k\mu}))^{-2\mu R/v-4}$. Therefore, $\underline{\lim}_{\lambda\to k\mu-}\frac{\partial \mathcal{W}(\lambda,d,v,k,R)}{\partial\lambda} > 0$. And part 2 is proved.

Part 3. Take derivative of $\mathcal{W}(\lambda, d, v, k, R)$ with respect to λ, we have that

$$\frac{\partial \mathcal{W}(\lambda,d,v,k,R)}{\partial\lambda} = \frac{\sum_{i=2}^{\infty}\frac{i(i-1)\lambda^{i-2}}{\mu_1\mu_2\cdots\mu_i}(1+\sum_{i=1}^{\infty}\frac{\lambda^i}{\mu_1\mu_2\cdots\mu_i}) - (\sum_{i=1}^{\infty}\frac{i\lambda^{i-1}}{\mu_1\mu_2\cdots\mu_i})^2}{(1+\sum_{i=1}^{\infty}\frac{\lambda^i}{\mu_1\mu_2\cdots\mu_i})^2}.$$

Let $\mu = v/d$, $a_i(\mu) = \frac{\lambda^i}{\mu_1\mu_2\cdots\mu_i}$, $b_i(\mu) = \frac{i\lambda^{i-1}}{\mu_1\mu_2\cdots\mu_i}$ and $c_i(\mu) = \frac{i(i-1)\lambda^{i-2}}{\mu_1\mu_2\cdots\mu_i}$. We first show

that for any μ, k, R and $0 \le \lambda < k\mu$, $\sum_{i=1}^{\infty} a_i(\mu)$, $\sum_{i=1}^{\infty} b_i(\mu)$ and $\sum_{i=2}^{\infty} c_i(\mu)$ are all finite. To do so, define $\epsilon = \frac{1}{2}(k/\lambda - 1/\mu) > 0$ and $N = \min\{i : 0 \le \frac{R}{(i-k+1)v} \le \epsilon\}$. We have

$$\begin{aligned}
\sum_{i=1}^{\infty} a_i(\mu) &= \sum_{i=1}^{N} \frac{\lambda^i}{\mu_1\mu_2\cdots\mu_i} + \frac{\lambda^N}{\mu_1\mu_2\cdots\mu_N} \sum_{i=1}^{\infty} \frac{\lambda^i}{\mu_{N+1}\mu_{N+2}\cdots\mu_{N+i}} \\
&< \sum_{i=1}^{N} \frac{\lambda^i}{\mu_1\mu_2\cdots\mu_i} + \frac{\lambda^N}{\mu_1\mu_2\cdots\mu_N} \sum_{i=1}^{\infty} \frac{\lambda^i(1/\mu+\epsilon)^i}{k^i} \\
&< \sum_{i=1}^{N} \frac{\lambda^i}{\mu_1\mu_2\cdots\mu_i} + \frac{\lambda^N}{\mu_1\mu_2\cdots\mu_N} \frac{k\mu+\lambda}{k\mu-\lambda} < \infty,
\end{aligned}$$

where the first inequality is because $\mu_l > \frac{k}{1/\mu+\epsilon}$ for any $l \ge N+1$. Similarly, we can also prove $\sum_{i=1}^{\infty} b_i < \infty$, and $\sum_{i=2}^{\infty} c_i < \infty$ when $0 \le \lambda < k\mu$.

With the finiteness of $\sum_{i=1}^{\infty} a_i$, $\sum_{i=1}^{\infty} b_i$ and $\sum_{i=2}^{\infty} c_i$, $\frac{\partial \mathcal{W}(\lambda,d,v,k,R)}{\partial \lambda}$ is finite as well. We also note that $a_i(\mu), b_i(\mu)$ and $c_i(\mu)$ are all continuous and decreasing in μ. Therefore by the monotone convergence theorem,

$$\lim_{\mu\to\infty} \sum_{i=1}^{\infty} a_i(\mu) = \sum_{i=1}^{\infty} \lim_{\mu\to\infty} a_i(\mu), \lim_{\mu\to\infty} \sum_{i=1}^{\infty} b_i(\mu) = \sum_{i=1}^{\infty} \lim_{\mu\to\infty} b_i(\mu),$$

and $\lim_{\mu\to\infty} \sum_{i=2}^{\infty} c_i(\mu) = \sum_{i=2}^{\infty} \lim_{\mu\to\infty} c_i(\mu)$. We thus have that for any given k, R and $0 \le \lambda < k\mu$,

$$\begin{aligned}
&\lim_{\mu\to\infty} \frac{\partial \mathcal{W}}{\partial \lambda} \\
= \;& \frac{\lim_{\mu\to\infty} \sum_{i=2}^{\infty} \frac{i(i-1)\lambda^{i-2}}{\mu_1\mu_2\cdots\mu_i} (1 + \lim_{\mu\to\infty} \sum_{i=1}^{\infty} \frac{\lambda^i}{\mu_1\mu_2\cdots\mu_i}) - \lim_{\mu\to\infty} (\sum_{i=1}^{\infty} \frac{i\lambda^{i-1}}{\mu_1\mu_2\cdots\mu_i})^2}{\lim_{\mu\to\infty} (1 + \sum_{i=1}^{\infty} \frac{\lambda^i}{\mu_1\mu_2\cdots\mu_i})^2} \\
= \;& \frac{\sum_{i=2}^{\infty} \frac{i(i-1)\lambda^{i-2}}{\lim_{\mu\to\infty}(\mu_1\mu_2\cdots\mu_i)} (1 + \sum_{i=1}^{\infty} \frac{\lambda^i}{\lim_{\mu\to\infty}(\mu_1\mu_2\cdots\mu_i)}) - (\sum_{i=1}^{\infty} \frac{i\lambda^{i-1}}{\lim_{\mu\to\infty}(\mu_1\mu_2\cdots\mu_i)})^2}{(1 + \sum_{i=1}^{\infty} \frac{\lambda^i}{\lim_{\mu\to\infty}(\mu_1\mu_2\cdots\mu_i)})^2} \\
= \;& \frac{\partial \overline{\mathcal{W}}}{\partial \lambda},
\end{aligned}$$

where $\overline{\mathcal{W}}$ is defined in Lemma A.1.1. Here the first equation holds since all limits exist and are finite, and the limit for the denominator is not zero.

By Lemma A.1.1, for any given R, v and k, there exists $\lambda^*(v,k,R)$ such that $\frac{\partial\overline{\mathcal{W}}}{\partial\lambda}\Big|_{\lambda=\lambda^*(v,k,R)} < 0$. Thus, by (A.4), there exists μ^* such that when $\mu = v/d > \mu^*$, that is, when $d < v/\mu^*$, we have that $\frac{\partial\mathcal{W}}{\partial\lambda}\big|_{\lambda=\lambda^*(v,k,R)} < 0$.

Part 4. By standard queueing theory, $\mathcal{W}(\lambda,d,v,k,R) = \frac{\sum_{i=1}^{\infty}\frac{i\lambda^{i-1}}{\mu_1\mu_2\cdots\mu_i}}{1+\sum_{i=1}^{\infty}\frac{\lambda^i}{\mu_1\mu_2\cdots\mu_i}} - d/v$. Now we consider a new approximated one-direction call system with arrival rate $\tilde{\lambda} = R\lambda$, average ride distance $\tilde{d} = d/R$, travel speed $\tilde{v} = v$, taxi number $\tilde{k} = k$ and road length $\tilde{R} = 1$. Note that the new system is equivalent to the original system after scaling the time in the original system by a factor of $1/R$ (one unit of time in the original system corresponds to $1/R$ units of time in the new system). Therefore the corresponding expected waiting time satisfies that $\mathcal{W}(\tilde{\lambda},\tilde{d},\tilde{v},\tilde{k},\tilde{R}) = \mathcal{W}(R\lambda,d/R,v,k,1) = \frac{1}{R}\mathcal{W}(\lambda,d,v,k,R)$. We have

$$\frac{\partial\mathcal{W}(\lambda,d,v,k,R)}{\partial\lambda} = R\frac{\partial\mathcal{W}(R\lambda,d/R,v,k,1)}{\partial\lambda} = R^2\frac{\partial\mathcal{W}(\tilde{\lambda},\tilde{d},\tilde{v},\tilde{k},\tilde{R})}{\partial\tilde{\lambda}}.$$

To prove that for any given d, v and $k \geq 2$, there exists $R^*(d,v,k)$ such that when $R > R^*(d,v,k)$, there exists $0 \leq \lambda(d,v,k,R) < kv/d$ satisfying $\frac{\partial\mathcal{W}(\lambda,d,v,k,R)}{\partial\lambda}\Big|_{\lambda=\lambda(d,v,k,R)} < 0$, it is equivalent to show that for the new approximated system, for any given d, v and $k \geq 2$, there exists a constant $R^*(d,v,k)$ such that when $R > R^*(d,v,k)$, there exists $0 \leq \lambda(d,v,k,R) < kv/d$ such that $\frac{\partial\mathcal{W}(\tilde{\lambda},\tilde{d},\tilde{v},\tilde{k},\tilde{R})}{\partial\tilde{\lambda}}\Big|_{\tilde{\lambda}=R\lambda(d,v,k,R)} < 0$.

By applying the result in Part 3, for the new approximated system, there exists an arrival rate $0 \leq \tilde{\lambda}^*(\tilde{v},\tilde{k},\tilde{R}) < \tilde{k}\tilde{v}/\tilde{d}$ and $\tilde{\mu}^*(\tilde{\lambda}^*,\tilde{v},\tilde{k},\tilde{R})$ such that $\frac{\partial\mathcal{W}(\tilde{\lambda},\tilde{d},\tilde{v},\tilde{k},\tilde{R})}{\partial\tilde{\lambda}}\Big|_{\tilde{\lambda}=\tilde{\lambda}^*(\tilde{v},\tilde{k},\tilde{R})} < 0$ for any $\tilde{d} < \tilde{v}/\tilde{\mu}^*(\tilde{\lambda}^*,\tilde{v},\tilde{k},\tilde{R})$. Since $\tilde{d} = d/R$, $\tilde{v} = v$, $\tilde{k} = k$ and $\tilde{R} = 1$, equivalently, there exists an arrival rate $0 \leq \tilde{\lambda}^*(v,k,1) < kRv/d$ such that $\frac{\partial\mathcal{W}(\tilde{\lambda},\tilde{d},\tilde{v},\tilde{k},\tilde{R})}{\partial\tilde{\lambda}}\Big|_{\tilde{\lambda}=\tilde{\lambda}^*(v,k,1)} < 0$ when $d/R < v/\tilde{\mu}^*(\tilde{\lambda}^*(v,k,1),v,k,1)$. Note that here $\tilde{\lambda}^*(v,k,1)$ only depends on v and k.

For given d, v and k, when choosing $R > R^*(d,v,k) = \max\left\{\frac{d\tilde{\mu}^*(\tilde{\lambda}^*(v,k,1),v,k,1)}{v}, \frac{d\tilde{\lambda}^*(v,k,1)}{kv}\right\}$, inequalities $\tilde{\lambda}^*(v,k,1) < kRv/d$ and $d/R < v/\tilde{\mu}^*(\tilde{\lambda}^*(v,k,1),v,k,1)$ are both satisfied. If we further choose $\lambda(v,k,R)$ by $\lambda(d,v,k,R) = \tilde{\lambda}^*(v,k,1)/R$, combining with $R > R^*(d,v,k)$, we have $\frac{\partial \mathcal{W}(\tilde{\lambda},\tilde{d},\tilde{v},\tilde{k},\tilde{R})}{\partial \tilde{\lambda}}\Big|_{\tilde{\lambda}=R\lambda(d,v,k,R)} < 0$.

Part 5. Consider a new approximated one-direction call system with travel speed $\tilde{v} = \alpha v$, and the same arrival rate $\tilde{\lambda} = \lambda$, average ride distance $\tilde{d} = d$, taxi number $\tilde{k} = k$ and road length $\tilde{R} = R$. Let $\widetilde{\mu_n}$ and $\widetilde{\mathcal{W}}(\widetilde{\lambda},\widetilde{d},\widetilde{v},\widetilde{k},\widetilde{R})$ denote the corresponding state dependent service rate and the average waiting time correspondingly. Since

$$\mu_n = \begin{cases} \frac{n}{\frac{d}{v}+\frac{R}{v(k-n+2)}}, & \text{if } n \le k \\ \frac{k}{\frac{d}{v}+\frac{R}{v(n-k+1)}}, & \text{if } n > k \end{cases} \tag{A.4}$$

it is easy to see that $\widetilde{\mu_n} = \alpha\mu_n$. As a result,

$$\begin{aligned} \widetilde{\mathcal{W}}(\widetilde{\lambda},\widetilde{d},\widetilde{v},\widetilde{k},\widetilde{R}) &= \frac{\sum_{i=1}^{\infty}\frac{i\widetilde{\lambda}^{i-1}}{\widetilde{\mu_1}\widetilde{\mu_2}\cdots\widetilde{\mu_i}}}{1+\sum_{i=1}^{\infty}\frac{\widetilde{\lambda}^i}{\widetilde{\mu_1}\widetilde{\mu_2}\cdots\widetilde{\mu_i}}} - \widetilde{d}/\widetilde{v} \\ &= \frac{1}{\alpha}\frac{\sum_{i=1}^{\infty}\frac{i(\lambda/\alpha)^{i-1}}{\mu_1\mu_2\cdots\mu_i}}{1+\sum_{i=1}^{\infty}\frac{(\lambda/\alpha)^i}{\mu_1\mu_2\cdots\mu_i}} - \frac{d}{\alpha v} \\ &= \frac{1}{\alpha}\mathcal{W}(\lambda/\alpha,d,v,k,R). \end{aligned}$$

Note that $\frac{\partial\widetilde{\mathcal{W}}(\widetilde{\lambda},\widetilde{d},\widetilde{v},\widetilde{k},\widetilde{R})}{\partial\widetilde{\lambda}} = \frac{\partial\mathcal{W}(\lambda,d,\alpha v,k,R)}{\partial\lambda} = \frac{1}{\alpha^2}\frac{\partial\mathcal{W}(\hat{\lambda},d,v,k,R)}{\partial\hat{\lambda}}|_{\hat{\lambda}=\lambda/\alpha}$. Since $\mathcal{W}(\hat{\lambda},d,v,k,R)$ has the same monotonicity (non-monotonicity) property in terms of the arrival rate λ as the $\mathcal{W}(\lambda,d,v,k,R)$, we have that $\mathcal{W}(\lambda,d,\alpha v,k,R)$ has the same monotonicity (non-monotonicity) in terms of the arrival rate λ as the $\mathcal{W}(\lambda,d,v,k,R)$. Therefore, the monotonicity (non-monotonicity) property of the average waiting time $\mathcal{W}(\lambda,d,v,k,R)$ in terms of λ is consistent with that of $\mathcal{W}(\lambda,d,\alpha v,k,R)$. Thus the theorem is proved. □

A.3 Proof of Theorem 3.3.2

Proof of Theorem 3.3.2. By redefining $R' = R/2$ and comparing $\mu_n^{(1c)}$ and $\mu_n^{(2c)}$, we note that the approximated two-direction call system is identical with the approximated one-direction call system with $R' = R/2$. That is, $\mathcal{W}^{(2c)}(\lambda, d, v, k, R) = \mathcal{W}^{(1c)}(\lambda, d, v, k, R')$. Thus the theorem follows directly from the proof of Theorem 3.3.1. □

A.4 Proof of Theorem 3.3.3

Proof of Theorem 3.3.3. Note that

$$\mu_n^{(nc)} = \begin{cases} \frac{n}{d/v+\int_0^{R/v}\left(\frac{R-vt}{R}\right)^{k-n+1}\exp\left(-\frac{(n-1)v^2t^2}{2Rd}\right)dt}, & n \le k, \\ \frac{k}{d/v+\int_0^{R/v}\left(\frac{R-vt}{R}\right)^{n-k}\exp\left(-\frac{\lambda vt^2}{2R}\right)dt}, & n > k. \end{cases}$$

$$= \begin{cases} \frac{n}{d/v+\int_0^1 \frac{R}{v}(1-x)^{k-n+1}\exp\left(-\frac{(n-1)Rx^2}{2d}\right)dx}, & n \le k, \\ \frac{n}{d/v+\int_0^1 \frac{R}{v}(1-x)^{n-k}\exp\left(-\frac{\lambda Rx^2}{2v}\right)dx}, & n > k. \end{cases}$$

We have for any $R \ge 0$,

$$\min\{n,k\}\mu \ge \mu_n^{(nc)} \ge \begin{cases} \frac{n}{d/v+\int_0^1 \frac{R}{v}(1-x)^{k-n+1}\exp\left(-\frac{(n-1)Rx^2}{2d}\right)dx}, & 0 \le n \le k, \\ \frac{k}{d/v+\int_0^1 \frac{R}{v}(1-x)^{n-k}\exp\left(-\frac{\lambda Rx^2}{2v}\right)dx}, & n > k. \end{cases}$$

Thus, $\lim_{R\to 0}\mu_n^{(nc)} = \min\{n,k\}\mu$. That is, when R approaches zero, the one-direction no-call approximation approaches a standard $M/M/k$ queue with arrival rate λ and service rate $\min\{n,k\}\mu$, for which the average waiting time increases with arrival rate λ.

For waiting time, we have $\mathcal{W}^{(nc)} = \frac{\sum_{i=1}^{\infty}\frac{i\lambda^{i-1}}{\mu_1^{(nc)}\mu_2^{(nc)}\cdots\mu_i^{(nc)}}}{1+\sum_{i=1}^{\infty}\frac{\lambda^i}{\mu_1^{(nc)}\mu_2^{(nc)}\cdots\mu_i^{(nc)}}} - 1/\mu$. By the bounded convergence theorem, we have $\lim_{R\to 0}\frac{\partial\mathcal{W}^{(nc)}}{\partial\lambda} = \frac{\partial\lim_{R\to 0}\mathcal{W}^{(nc)}}{\partial\lambda} = \frac{\partial\mathcal{W}_{M/M/k}}{\partial\lambda} > 0$. Since $\mathcal{W}^{(nc)}$ is differentiable in λ, when R is small enough, we must have $\frac{\partial\mathcal{W}^{(nc)}}{\partial\lambda} > 0$. □

A.5 Proof of Theorem 3.4.1

Proof of Theorem 3.4.1. It is easy to see that $\mu_n^{(nc)} > \mu_n^{(1c)}$ for all n (because $\mu_n^{(1c)}$ is just $\mu_n^{(nc)}$ without the exponential part in the denominator). That is, the approximated one-direction no-call system has a higher service rate compared to the approximated one-direction call system for any system state. Now we show that this implies the average waiting time in the no-call system is shorter. To show that, we show that in a state-dependent $M/M/k$ queue, the average waiting time $\mathcal{W}(\lambda, \mu)$ is decreasing in μ_m for any m, i.e., $\frac{\partial \mathcal{W}}{\partial \mu_m} < 0$ for any m and μ_m. We have

$$\mathcal{W} = \frac{\sum_{i=1}^{\infty} \frac{i\lambda^{i-1}}{\mu_1\mu_2\cdots\mu_i}}{1+\sum_{i=1}^{\infty} \frac{\lambda^i}{\mu_1\mu_2\cdots\mu_i}} - 1/\mu = \frac{1}{\lambda}\frac{c_1+\sum_{i=m}^{\infty} \frac{i\lambda^i}{\mu_1\mu_2\cdots\mu_i}}{c_2+\sum_{i=m}^{\infty} \frac{\lambda^i}{\mu_1\mu_2\cdots\mu_i}} - 1/\mu,$$

where $c_1 = \sum_{i=1}^{m-1} \frac{i\lambda^i}{\mu_1\mu_2\cdots\mu_i}$ and $c_2 = 1+\sum_{i=1}^{m-1} \frac{\lambda^i}{\mu_1\mu_2\cdots\mu_i}$.

Define $f = \frac{c_1+\sum_{i=m}^{\infty} \frac{i\lambda^i}{\mu_1\mu_2\cdots\mu_i}}{c_2+\sum_{i=m}^{\infty} \frac{\lambda^i}{\mu_1\mu_2\cdots\mu_i}} = \frac{c_1+\sum_{i=m}^{\infty} \frac{a_i}{\mu_m}}{c_2+\sum_{i=m}^{\infty} \frac{b_i}{\mu_m}}$, where $a_i = \frac{i\lambda^i}{\mu_1\mu_2\cdots\mu_{m-1}\mu_{m+1}\cdots\mu_i}$ and $b_i = \frac{a_i}{i}$. To prove $\frac{\partial \mathcal{W}}{\partial \mu_m} < 0$, it is equivalent to prove that $\frac{\partial f}{\partial \mu_m} < 0$. We have that

$$\begin{aligned}\frac{\partial f}{\partial \mu_m} &= \frac{\sum_{i=m}^{\infty} \frac{b_i}{\mu_m^2}(c_1+\sum_{i=m}^{\infty} \frac{a_i}{\mu_m}) - (\sum_{i=m}^{\infty} \frac{a_i}{\mu_m^2})(c_2+\sum_{i=m}^{\infty} \frac{b_i}{\mu_m})}{(c_2+\sum_{i=m}^{\infty} \frac{b_i}{\mu_m})^2} \\ &= \frac{c_1\sum_{i=m}^{\infty} b_i - c_2\sum_{i=m}^{\infty} a_i}{\mu_m^2(c_2+\sum_{i=m}^{\infty} \frac{b_i}{\mu_m})^2}.\end{aligned}$$

Note that $c_1 < mc_2$ and $\sum_{i=m}^{\infty} a_i > m\sum_{i=m}^{\infty} b_i$, we thus get $c_1\sum_{i=m}^{\infty} b_i - c_2\sum_{i=m}^{\infty} a_i < 0$. As a result, $\frac{\partial f}{\partial \mu_m} < 0$ and $\frac{\partial \mathcal{W}}{\partial \mu_m} < 0$. Therefore, the average waiting time in the approximated one-direction no-call system is always smaller than that in the approximated one-direction call system, that is, $\mathcal{W}^{(nc)} < \mathcal{W}^{(1c)}$. □

A.6 Proof of Lemma A.6.1

Lemma A.6.1 *Let $\mu_n^{(2c)}$ and $\mu_n^{(nc)}$ denote the state dependent service rate in two-direction call and no-call systems respectively. We would have $\mu_n^{(2c)} \geq \mu_n^{(nc)}$ when $n \geq \sqrt{\frac{\lambda R}{v\log 4}} + k - 2$.*

Proof of Lemma A.6.1. Note that

$$\mu_n^{(nc)} = \begin{cases} \frac{n}{d/v + \int_0^{R/v} \left(\frac{R-vx}{R}\right)^{k-n+1} \exp\left(-\frac{(n-1)v^2x^2}{2Rd}\right) dx}, & n \leq k, \\ \frac{k}{d/v + \int_0^{R/v} \left(\frac{R-vx}{R}\right)^{n-k} \exp\left(-\frac{\lambda v x^2}{2R}\right) dx}, & n > k. \end{cases}$$

$$\mu_n^{(2c)} = \begin{cases} \frac{n}{d/v + \frac{R}{2(k-n+2)v}} = \frac{n}{d/v + \int_0^{R/v} \frac{1}{2}\left(\frac{R-vx}{R}\right)^{k-n+1} dx}, & \text{if } n \leq k, \\ \frac{k}{d/v + \frac{R}{2(n-k+1)v}} = \frac{k}{d/v + \int_0^{R/v} \frac{1}{2}\left(\frac{R-vx}{R}\right)^{n-k} dx}, & \text{if } n > k. \end{cases}$$

To show that $\mu_n^{(2c)} > \mu_n^{(nc)}$ when n is large enough is equivalent to show that $\int_0^{R/v} \left(\frac{R-vx}{R}\right)^{n-k} \exp\left(-\frac{\lambda v x^2}{2R}\right) dx > \int_0^{R/v} \frac{1}{2}\left(\frac{R-vx}{R}\right)^{n-k} dx$ when n is large enough. Note that $\frac{\int_0^{R/v} \left(\frac{R-vx}{R}\right)^{n-k} \exp\left(-\frac{\lambda v x^2}{2R}\right) dx}{\int_0^{R/v} \left(\frac{R-vx}{R}\right)^{n-k} dx} = \mathrm{E}\left[\exp\left(-\frac{\lambda v X^2}{2R}\right)\right]$, where X is a random variable with density function $f(x) = \frac{\left(\frac{R-vx}{R}\right)^{n-k}}{\int_0^{R/v} \left(\frac{R-vx}{R}\right)^{n-k} dx}$. Since $\exp\left(-\frac{\lambda v x^2}{2R}\right)$ is convex in x, by Jensen's inequality, we have that $\mathrm{E}\left[\exp\left(-\frac{\lambda v X^2}{2R}\right)\right] \geq \exp\left(-\frac{\lambda v \mathrm{E}[X]^2}{2R}\right) = \exp\left(-\frac{\lambda R}{2v(n-k+2)^2}\right)$. When $n \geq \sqrt{\frac{\lambda R}{v\log 4}} + k - 2$, we have $\exp\left(-\frac{\lambda R}{2v(n-k+2)^2}\right) \geq \frac{1}{2}$, resulting in

$$\int_0^{R/v} \left(\frac{R-vx}{R}\right)^{n-k} \exp\left(-\frac{\lambda v x^2}{2R}\right) dx > \int_0^{R/v} \frac{1}{2}\left(\frac{R-vx}{R}\right)^{n-k} dx.$$

Therefore, $\mu_n^{(2c)} \geq \mu_n^{(nc)}$ when $n \geq \sqrt{\frac{\lambda R}{v\log 4}} + k - 2$. □

A.7 Proof of Theorem 3.4.2

Proof of Theorem 3.4.2.

Part 1. We first prove that $\lim_{\lambda\to 0^+} \mathcal{W}^{(2c)} < \lim_{\lambda\to 0^+} \mathcal{W}^{(nc)}$. Note that when $\lambda \to 0+$, we have that $\lambda/\mu_i^{(2c)} \to 0$ and $\lambda/\mu_i^{(nc)} \to 0$ for any $i \in \mathbb{Z}_+$.

$$\begin{aligned}\lim_{\lambda\to 0^+} \mathcal{W}^{(nc)} &= \lim_{\lambda\to 0^+} \frac{\sum_{i=1}^{\infty} \frac{i\lambda^{i-1}}{\mu_1^{(nc)}\mu_2^{(nc)}\cdots\mu_i^{(nc)}}}{1+\sum_{i=1}^{\infty} \frac{\lambda^i}{\mu_1^{(nc)}\mu_2^{(nc)}\cdots\mu_i^{(nc)}}} - d/v \\ &= = 1/\mu_1^{(nc)} - d/v = \int_0^{R/v} \left(\frac{R-vx}{R}\right)^k dx.\end{aligned}$$

Similarly, we have

$$\begin{aligned}\lim_{\lambda\to 0^+} \mathcal{W}^{(2c)} &= \lim_{\lambda\to 0^+} \frac{\sum_{i=1}^{\infty} \frac{i\lambda^{i-1}}{\mu_1^{(2c)}\mu_2^{(2c)}\cdots\mu_i^{(2c)}}}{1+\sum_{i=1}^{\infty} \frac{\lambda^i}{\mu_1^{(2c)}\mu_2^{(2c)}\cdots\mu_i^{(2c)}}} - d/v \\ &= 1/\mu_1^{(2c)} - d/v = \frac{1}{2}\int_0^{R/v} \left(\frac{R-vx}{R}\right)^k dx.\end{aligned}$$

It is easy to see that

$$\lim_{\lambda\to 0^+} \mathcal{W}^{(2c)} = \frac{1}{2}\lim_{\lambda\to 0^+} \mathcal{W}^{(nc)} < \lim_{\lambda\to 0^+} \mathcal{W}^{(nc)}.$$

Part 2. Now we prove that $\lim_{\lambda\to kv/d^-} \mathcal{W}^{(2c)} < \lim_{\lambda\to kv/d^-} \mathcal{W}^{(nc)}$. First, we analyze the term $\lim_{\lambda\to kv/d^-} \mathcal{W}^{(2c)}$. Note that $\mu_n^{(2c)}$ monotonically increases in n when $n \geq k+1$ and it converges to kv/d as n goes to infinity. For any fixed arrival rate λ satisfying $\lambda \geq \mu_{k+1}^{(2c)}$, there exists an integer $n^*(\lambda)$ such that $\mu_{n^*}^{(2c)} \leq \lambda < \mu_{n^*+1}^{(2c)}$. We can see that $\frac{\lambda}{\mu_i^{(2c)}} \geq 1$ for $k+1 \leq i \leq n^*$, and $\frac{\lambda}{\mu_i^{(2c)}} < 1$ for any $i \geq n^*+1$. Define $b^* = \sqrt{\frac{\lambda R}{v\log 4}} + k - 2$, from Lemma A.6.1, it is known that $\mu_n^{(2c)} \geq \mu_n^{(nc)}$ when $n \geq b^*$. Since $\lambda < kv/d$,

we have $b^* < \sqrt{\frac{kR}{d\log 4}} + k - 2$. Note that n^* increases and approaches infinity as λ approaches kv/d. Therefore, when λ is close enough to kv/d, we have $n^* > b^*$. In $\mathcal{W}^{(2c)} = \frac{\sum_{i=1}^{\infty} \frac{i\lambda^{i-1}}{\mu_1^{(2c)}\mu_2^{(2c)}\cdots\mu_i^{(2c)}}}{1+\sum_{i=1}^{\infty} \frac{\lambda^i}{\mu_1^{(2c)}\mu_2^{(2c)}\cdots\mu_i^{(2c)}}} - d/v$, the numerator of the first term satisfies that

$$\sum_{i=1}^{\infty} \frac{i\lambda^{i-1}}{\mu_1^{(2c)}\mu_2^{(2c)}\cdots\mu_i^{(2c)}} = \sum_{i=1}^{b^*} \frac{i\lambda^{i-1}}{\mu_1^{(2c)}\mu_2^{(2c)}\cdots\mu_i^{(2c)}} + \sum_{i=b^*+1}^{\infty} \frac{i\lambda^{i-1}}{\mu_1^{(2c)}\mu_2^{(2c)}\cdots\mu_i^{(2c)}}.$$

Define $p = \frac{\lambda}{\mu_{n^*}^{(2c)}} > 1$ and $q = \frac{\lambda d}{kv} < 1$. We note that

$$\begin{aligned}
& \sum_{i=b^*+1}^{\infty} \frac{i\lambda^{i-1}}{\mu_1^{(2c)}\mu_2^{(2c)}\cdots\mu_i^{(2c)}} \\
= & \sum_{i=b^*+1}^{n^*} \frac{i\lambda^{i-1}}{\mu_1^{(2c)}\mu_2^{(2c)}\cdots\mu_i^{(2c)}} + \sum_{i=n^*+1}^{\infty} \frac{i\lambda^{i-1}}{\mu_1^{(2c)}\mu_2^{(2c)}\cdots\mu_i^{(2c)}} \\
\geq & \frac{\lambda^{b^*-1}}{\mu_1^{(2c)}\cdots\mu_{b^*}^{(2c)}} \left(\sum_{i=1}^{n^*-b^*} (b^*+i)p^i + \sum_{i=1}^{\infty} p^{n^*-b^*}(n^*+i)q^i \right) \\
= & \frac{\lambda^{b^*-1}}{\mu_1^{(2c)}\cdots\mu_{b^*}^{(2c)}} \left(\frac{(n^*(p-1)-1)p^{n^*-b^*+1} - b^*p^2 + (b^*+1)p}{(p-1)^2} \right. \\
& \left. + p^{n^*-b^*} \frac{q(1-q)n^*+q}{(1-q)^2} \right) \\
= & \mathcal{O}(p^{n^*})
\end{aligned} \tag{A.5}$$

When n^* is large enough, we have $\sum_{i=1}^{b^*} \frac{i\lambda^{i-1}}{\mu_1^{(2c)}\mu_2^{(2c)}\cdots\mu_i^{(2c)}} \ll \sum_{i=b^*+1}^{\infty} \frac{i\lambda^{i-1}}{\mu_1^{(2c)}\mu_2^{(2c)}\cdots\mu_i^{(2c)}}$. Following similar analysis, we can see that the denominator of the first term in $\mathcal{W}^{(2c)}$

satisfies that $1+\sum_{i=1}^{b}\frac{\lambda^i}{\mu_1^{(2c)}\mu_2^{(2c)}\cdots\mu_i^{(2c)}} \ll \sum_{i=b^*+1}^{\infty}\frac{\lambda^{i-1}}{\mu_1^{(2c)}\mu_2^{(2c)}\cdots\mu_i^{(2c)}}$. As a result,

$$\begin{aligned}
\mathcal{W}^{(2c)} &= \frac{\sum_{i=1}^{b^*}\frac{i\lambda^{i-1}}{\mu_1^{(2c)}\mu_2^{(2c)}\cdots\mu_i^{(2c)}}+\sum_{i=b^*+1}^{\infty}\frac{i\lambda^{i-1}}{\mu_1^{(2c)}\mu_2^{(2c)}\cdots\mu_i^{(2c)}}}{1+\sum_{i=1}^{b^*}\frac{\lambda^i}{\mu_1^{(2c)}\mu_2^{(2c)}\cdots\mu_i^{(2c)}}+\sum_{i=b^*+1}^{\infty}\frac{\lambda^i}{\mu_1^{(2c)}\mu_2^{(2c)}\cdots\mu_i^{(2c)}}} - d/v \\
&= \frac{\sum_{i=b^*+1}^{\infty}\frac{i\lambda^{i-1}}{\mu_1^{(2c)}\mu_2^{(2c)}\cdots\mu_i^{(2c)}}}{\sum_{i=b^*+1}^{\infty}\frac{\lambda^i}{\mu_1^{(2c)}\mu_2^{(2c)}\cdots\mu_i^{(2c)}}} + o(1) - d/v \\
&= \frac{\sum_{i=1}^{\infty}\frac{(b^*+i)\lambda^{i-1}}{\mu_{b^*+1}^{(2c)}\mu_{b^*+2}^{(2c)}\cdots\mu_{b^*+i}^{(2c)}}}{\sum_{i=1}^{\infty}\frac{\lambda^i}{\mu_{b^*+1}^{(2c)}\mu_{b^*+2}^{(2c)}\cdots\mu_{b^*+i}^{(2c)}}} - d/v + o(1) \\
&= \frac{\sum_{i=1}^{\infty}\frac{i\lambda^{i-1}}{\mu_{b^*+1}^{(2c)}\mu_{b^*+2}^{(2c)}\cdots\mu_{b^*+i}^{(2c)}}}{\sum_{i=1}^{\infty}\frac{\lambda^i}{\mu_{b^*+1}^{(2c)}\mu_{b^*+2}^{(2c)}\cdots\mu_{b^*+i}^{(2c)}}} + \frac{b^*}{\lambda} - d/v + o(1)
\end{aligned}$$

In a similar manner, we obtain that

$$\mathcal{W}^{(nc)} = \frac{\sum_{i=1}^{\infty}\frac{i\lambda^{i-1}}{\mu_{b^*+1}^{(nc)}\mu_{b^*+2}^{(nc)}\cdots\mu_{b^*+i}^{(nc)}}}{\sum_{i=1}^{\infty}\frac{\lambda^i}{\mu_{b^*+1}^{(nc)}\mu_{b^*+2}^{(nc)}\cdots\mu_{b^*+i}^{(nc)}}} + \frac{b^*}{\lambda} - d/v + o(1). \tag{A.6}$$

Since $\mu_{b^*+i}^{(2c)} > \mu_{b^*+i}^{(nc)}$ for any $i \in \mathbb{Z}_+$, following a similar approach as in Theorem 3.4.1, we thus get $\mathcal{W}^{(nc)} > \mathcal{W}^{(2c)}$ when n^* is large enough, that is, when λ approaches kv/d. Therefore, the theorem is proved. □